UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Modeling Alzheimer's Disease progression using Temporal Data

Mafalda Cardoso de Lemos Gomes Beleza

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Professora Doutora Helena Isabel Aidos Lopes Tomás

2020

Agradecimentos

Os meus primeiros agradecimentos vão para a minha família por todo o apoio que me deram ao longo da realização da tese. Sem eles certamente não o teria sido possivel.

Quero também agradecer aos meus amigos de mestrado que acompanharam esta caminhada desde o início. E claro, a todos os meus amigos que acompanharam este caminho de fora, mas que em todos os momentos dificeis estiveram ao meu lado.

Deixo também um agradecimento especial à professora Helena Aidos não só pela orientação do trabalho, mas também por todo o apoio e paciência ao longo deste trabalho.

Abstract

Alzheimer's Disease (AD) is a progressive brain disorder that slowly leads to memory loss, confusion, disorientation, and inability to communicate. It is very important an early detection of the disease in order to improve patients' life quality and slow down the symptoms. Since there is still no cure available (although specific medications may attenuate the symptoms for a time), it ultimately draws from family members and society. Measuring and estimating the progression of such disease is therefore very important from both the medic and family's perspective.

Several studies have been made to address problems such as Alzheimer's disease diagnosis and prognosis by discovering biomarkers. However, only a few studies use temporal information to model disease progression patterns. Hence, the objective of this thesis is to model the progression patterns of the disease through neuropsychological tests, leading to a better understanding of the underlying disease mechanisms and improving prognosis. In that sense, several summarization and representation techniques were applied to the dataset composed by neuropsychological tests, and the performance of classification techniques were assessed. Experimental results showed that representation techniques, such as ESiG, have a higher sensitivity and specificity values than others summarization and representation techniques or even the static data, using one medical appointment to predict the progression of Alzheimer's disease.

Keywords: Alzheimer's Disease, Temporal Data, neuropsychological tests, time-series summarization techniques, time-series representation techniques.

Resumo

A doença de Alzheimer é uma doença progressiva no cérebro que lentamente leva a perda de memória, confusão, desorientação e incapacidade de comunicar. É muito importante a deteção precoce da doença para ser possível aumentar a qualidade de vida a abrandar os sintomas. Atualmente ainda não existe uma cura para esta doença (apesar de medicação especifica pode atenuar os sintomas por um tempo) em última análise, muito desejado por membros da família e sociedade. Quantificar e estimar a progressão de uma doença como esta é muito importante na perspetiva médica e familiar.

Diversos estudos foram realizados na abordagem de problemas como o diagnóstico da doença de Alzheimer e prognósticos por descoberta de biomarcadores. Contudo, apenas alguns estudos usam informação temporal para modelar os padrões de progressão da doença. Consequentemente, o objetivo desta tese é modelar padrões de progressão da doença com testes neuropsicológicos, para conduzir uma melhor compreensão dos mecanismos subjacentes da doença e melhorar o prognóstico. Neste Sentido, foram aplicadas diversas técnicas de sumarização e de representação ao conjunto de dados composto por testes neuropsicológicos, e avaliado o desempenho dos classificadores. Os resultados experimentais mostram que técnicas de representação, tais como o ESiG, apresentam valores de sensibilidade e de especificidade maiores que outras técnicas de summarização e de representação, ou mesmo utilizando valores de apenas uma consulta médica para prever a progressão para doença de Alzheimer.

Palavras-chaves: Doença de Alzheimer, Dados Temporais, Testes Neuropsicologicos, Técnicas de sumarização de dados temporais, técnicas de representação de dados temporais

Resumo alargado

A Doença de Alzheimer (DA) é uma doença neurodegenerativa que tem como principal impacto a deficiência cognitiva e alteração no comportamento. Esta doença é a forma de demência mais comum em todo o mundo, representando cerca de 60 a 80%. Atualmente atinge cerca de 45 milhões de pessoas em todo o mundo, e a previsão é que estes números aumentem para 75 milhões até 2030 e que triplique até 2050. Esta é uma doença relacionada com a idade, sendo que cerca de 3% da população com mais de 65 anos é atingida e com mais de 85 anos a prevalência aumenta para cerca de 30%. Existe um estado inicial da doença pré-demência, a deficiência cognitiva ligeira (DCL) e sabe-se que pacientes diagnosticados com DCL, têm um risco mais elevado de desenvolverem a DA. Esta doença tem como principais consequências a desorientação global, perda de funções cognitivas de forma progressiva e irreversível entre as quais está a memória, concentração, linguagem e pensamento. Esta perda de capacidades leva a uma incapacidade de realizar atividades da vida diária. Na origem desta doença está a formação de placas de proteína beta-amiloide nas células cerebrais. Esta acumulação e formação de placas impede a comunicação entre as células nervosas no cérebro provocando a morte celular. O primeiro diagnóstico criado para a DA foi estabelecido pelo instituto nacional do envelhecimento e associação de Alzheimer. Esta associação recomenda a realização de exames de imagem de Ressonância Magnética (RM) para quantificar o volume intracraniano, tomografia por emissão de positrões (TEP) para detetar alterações metabólicas, análise do fluido cefalorraquidiano para quantificação de biomarcadores e testes neuropsicológicos para medir o comprometimento cognitivo. Um dos maiores problemas da deteção desta doença é que normalmente é feito um diagnóstico tardio. Isto acontece porque os primeiros sintomas aparecem cerca de cinco anos após as primeiras alterações biológicas acontecerem.

Uma vez que doentes identificados com DCL têm elevado risco de desenvolverem a DA, torna-se importante distinguir os doentes DCL que irão desenvolver DA daqueles que permaneceram estáveis. A identificação deste doentes é extremamente útil para estudar os processos que ocorrem na progressão da DA. Neste contexto torna-se importante fazer alguns testes a pacientes DCL regularmente que não sejam dispendiosos, sem radiação, e que permita avaliar a progressão da doença, tais como os testes neuropsicológicos.

Existe um grande número de testes neuropsicológicos que devem ser realizados. Alguns dos exames mais frequentemente utilizados são o Exame Breve de Estado Mental (Mini-Mental State Examination - MMSE), que avalia a capacidade de leitura, escrita, orientação e memória a curto prazo; Escala de Avaliação para a Doença de Alzheimer – Cognitiva (Alzheimer Disease Assessment Scale – Congnitive – ADAS-Cog) que avalia a memória e capacidade linguística.

Para a realização desta tese utilizou-se dados da Cognitive Complaints Cohorts (CCC). Este conjunto de dados pretende estudar a progressão de demência em indivíduos com queixas cognitivas baseado na avaliação extensiva neuropsicológica em uma das seguintes instituições - Laboratório de Estudo da linguagem, no Hospital Santa Maria, ou Clinica da Memória, também em Lisboa. Para ser admitido no CCC os critérios de inclusão tiveram de ser todos cumpridos, nomeadamente ter queixa cognitivas

e completar a avaliação com bateria de testes para avaliar diversos domínios cognitivos validados para a população portuguesa (Bateria de Lisboa para Avaliação das Demenências (BLAD)). Foram também considerados critérios de exclusão tais como o diagnóstico de demência ou outros problemas de saúde que podem causar comprometimento cognitivo, tais como acidente vascular cerebral, tumor cerebral, trauma craniano significativo, Epilepsia, entre outros.

Na realização deste trabalho foram utilizadas diversas ferramentas de Aprendizagem Automática para conseguir analisar e extrair a informação útil dos dados. Neste trabalho iniciou-se uma análise com dados com informação dos primeiros exames neuropsicológicos realizados, dados estáticos. Com esta análise inicial pretende-se estabelecer um ponto de comparação com os resultados de dados temporais. Foram analisadas inicialmente o desempenho de quatro classificadores para distinguirem o tipo de evolução de cada indivíduo. Para esta classificação inicial utilizou-se os seguintes classificadores: Naive-Bayes, Decision Tree, Support Vector Machines e Redes Neuronais. Numa primeira classificação não houve nenhum classificador que mostrasse um desempenho muito acima dos restantes. Optámos por utilizar para uma primeira análise o classificador com melhor accuracy neste conjunto de dados, sendo por isto escolhido o classificador Support Vector Machine (SVM), que obteve o valor de 0.786 de accuracy.

Numa segunda fase da tese, foi tido em consideração o fator tempo. Para isto foram construídos novos conjuntos de dados que compreendiam a informação dos três primeiros exames neuropsicológicos para cada individuo. Nesta abordagem foram também construídas diferentes classes utilizando a abordagem de Time-Window, tendo sido construídos com base na informação após a terceira avaliação, e foram utilizadas janelas de tempo de um, dois e três anos. Uma vez que as duas primeiras janelas de tempo resultavam num conjunto de dados muito desbalanceado, o que levaria a demasiados exemplos simulados por técnicas de rebalanceamento, foi utilizado o conjunto de dados com as classes geradas por uma janela de tempo de três anos. O conjunto de dados escolhido continha uma grande quantidade de dados em falta, característica de conjunto de dados reais. Para colmatar a informação em falta neste conjunto de dados foram utilizadas duas abordagens diferentes. Inicialmente utilizada o preenchimento de dados em falta com a média dos valores para aquela observação. Uma segunda abordagem considera que no caso de não existir informação em contrário a avaliação permanece igual à anterior, e no caso desta não existir está no ponto zero. Com estes novos conjuntos de dados utilizamos técnicas de sumarização e de representação para representar a sequência de dados obtida e realizámos nova classificação com o classificador previamente selecionado, SVM. As técnicas de sumarização utilizadas foram a média e a mediana para criar conjuntos de dados que mimetizasse a informação do conjunto de dados originais. Com os conjuntos de dados relativamente à media o classificador apresentou uma accuracy de 0.785, sensibilidade de 0.835 e especificidade de 0.615. Relativamente à mediana os resultados obtidos de accuracy, sensibilidade e especificidade foram de 0.815, 0.866 e 0.725 respetivamente. As técnicas de representação utilizadas foram Discrete Wavelet Transform (DWT), ESiG, Symbolic Aggregate Approximation (SAX). Estas técnicas utilizam os dados sequenciais para construir novos conjuntos de dados representativos dos dados originais. Com os dados resultantes da transformação com DWT os

resultados de accuracy, sensibilidade e especificidade foram, 0.806, 0.904 e 0.439, respetivamente. Com a representação com ESiG os resultados de accuracy, sensibilidade e especificidade foram, 0.756, 0.881 e 0.959, respetivamente. Relativamente ao SAX os resultados de accuracy, sensibilidade e especificidade foram, 0.454, 0.947 e 0.053, respetivamente. Estes últimos resultados levaram a novas experiências com uma técnica que inclui o SAX e o classificador Vector Space Model associado, SAX-VSM. Com esta nova abordagem os resultados foram significativamente melhores relativamente à classificação com SVM, tendo sido de 0.600, 0.730 e 0.20, respetivamente para accuracy, sensibilidade e especificidade. Numa segunda abordagem com dados temporais tendo em consideração o segundo método de lidar com dados em falta, os resultados obtidos foram melhores para métodos de representação utilizando SAX, com accuracy de 0.775, sensibilidade 0.961 e de especificidade de 0.219 e SAX-VSM com valores de 0.821, 0.9167 e 0.333 respetivamente. Os outros dois tipos de representação obtiveram resultados ligeiramente inferiores.

Com estes resultados conseguimos perceber que o tratamento dos dados em falta numa sequência de dados pode alterar a maneira como diferentes técnicas de representação lidam com os dados e a informação útil que conseguem extrair da sequência. As técnicas de sumarização e de representação apresentaram resultados superiores de accuracy relativamente aos dados estáticos, mas com métricas de sensibilidade consideravelmente melhores. Mostrando assim ser uma mais valia considerar mais do que um ponto de avaliação neuropsicológica para um possível prognóstico antecipado.

Contents

List of Figures

List of Tables

Acronyms

AD Azheimer Disease

ADNI Alzheimer's Disease Neuroimaging Initiative

CCC Cognitive Complaints Cohort

CSF Cerebrospinal fluid

DFT Discrete Fourier Transform

DT Decision Tree

FN False Negative

FP False Positive

FS Feature Selection

ML Machine Learning

MCI Mild Cognitive Impairment

MMSE Mini-Mental State Examination

cMCI converted Mild Cognitive Impairment

sMCI stable Mild Cognitive Impairment

MSE Mean Square Error

MRI Magnetic Resonance Imaging

pre-MCI pre-Mild Cognitive Impairment

NN Neural Networks

PAA Piecewise Aggregate Approximation

PET Positron emission tomography

SAX Symbolic Aggregate Approximation

SAX-VSM Symbolic Aggregate Approximation - Vector Space Model

SMOTE Synthetic Minority Over-Sampling Technique

SVM Support Vector Machine

TN True Negative

TP True Positive

Chapter 1

Introduction

Currently, dementia affects more than 45 million people worldwide, and the most common cause of dementia is Alzheimer's disease (AD), around 60-80%. This number is expected to rise to 75 million people globally by 2030 and is estimated to triple by 2050[49]. Alzheimer's disease is a neurodegenerative disorder that affects cognition function and behavior. This disease is characterized by a long preclinical, mild cognitive impairment (MCI), and prodromal phase, around to twenty years before the diagnose, and has a clinical duration of approximately eight to ten years. In this disease occur changes that are indistinguishable for the patients along a lot of time. Alzheimer's is an age-related disease and occur in persons that have more than 65 years old, with a prevalence of 3%, and in persons that have more than 85 years or more, the prevalence goes up to 30%[16, 42, 3].

The amyloid cascade hypothesis has been accepted as an explanation of the disease. This hypothesis consists of the deposition of the amyloid-beta peptide in the brain that causes the toxic effect of neurofibrillary tangles, cell loss, and vascular damage, which, in turn, could culminate in dementia. However, autopsies on known AD patients revealed clusters outside neurons of beta-amyloid protein (plaques) and accumulation of an abnormal form of tau protein inside neurons (tangles)[22]. AD may be familial early-onset or sporadic late-onset, and these two forms are different. The early-onset AD is related to three genes autosomal dominant. The three genes include the amyloid precursor protein gene on chromosome 21, the presenilin 1 gene on chromosome 4, and the presenilin 2 gene in chromosome 1. The mutations on this gene lead to the overproduction of beta-amyloid protein, which gives rise to synaptic disfunction, neurotoxic and beta-amyloid deposits in brain cells. The presenilin 1 gene is the most common cause of early-onset AD. However, this form of AD is rare. In sporadic or late-onset AD, the apolipoprotein-E (APOE) e4 allele increase the risk of developing the disease. The gene may be involved in the degradation or clearance of the beta-amyloid protein from the brain. Other risks associated with AD also exist, such as having diabetes type 2, smoking, hypertension, sedentary lifestyle, obesity, and head injuries.

It does not yet exist a treatment available for this disease. One of the reasons for many clinical trials has been failed is that the diagnosis is made too late, and the disease processes already occur. However, there exists medication that reduces the symptoms of the disease[7]. As a consequence of this process,

cognitive function is compromised. Approximately five years after the beginning of processes, the first symptoms appear, like cognitive and functional decline, and memory lapses happen, difficulty to say everyday words, and other problems related to cognitive domains such as memory, attention, and language. With the advanced cognitive decline, the advanced stage of AD leads to total dependence of caregivers or families because patients become incapable of doing daily activities. In this context, early diagnosis has the most importance and impact on patients' quality of life. The early diagnosis allows the doctor to intervene in disease much earlier and offer the patient more time with independent life. The importance of this disease in the world led to creating an association to investigate the cure or modify the therapies that already exist. As a result, dementia research has been successfully carried out, notably in terms of clinical data collection, thus allowing for the most advanced diagnosis study and prognosis. The first diagnostic criteria for AD dementia, established in 1984, were revised in 2011 by the national institute of aging and the Alzheimer's Association. This association recommends exams like Magnetic Resonance Imaging (MRI) for quantifying the hippocampal volume and total intracranial volume, Positron emission tomography (PET), for recognition metabolic changes, cerebrospinal fluid (CSF) analysis to quantify biomarkers concentration, and neuropsychological testing for measure cognitive impairment, for diagnosing AD. In the criteria for dementia is required that the symptoms, cognitive and behavioral, cannot be explained by any other medical, neurologic, or psychiatric condition; there are functional impairments; and the behavioral or cognitive impairment involves at least two of the following domains(memory; executive functions; visuospatial abilities; language; and behavior, personality, or comportment)[49]. One of the major problems with Alzheimer's disease is a late diagnosis. This occurs since the symptoms of biological changes appear around five years after the beginning of biological changes. These changes have a direct impact on cognitive activity and are directly related to Alzheimer's disease. Many studies say that individuals with MCI have had higher risk conversion rates to dementia, particularly to AD. Thus, it is crucial to distinguish MCI subjects that will convert into AD and those that will remain stable. This distinction is the relevant factor that allows us to know those patients that should be studied to discover processes that occur in disease progress. In this context, it is appropriate to produce a prognosis of AD with the same exams performed in patients regularly without higher cost, without radiation, and evaluate the disease's progress, such as neuropsychological tests.

1.1 Objectives and Contributions

The main goal of this work is to automatically predict which MCI subjects will convert to Alzheimer's disease. In neurodegenerative diseases, it is extremely important to predict as earlier as possible the disease and know-how will progress in the future. Individuals with MCI can stay stable or progress to dementia like Alzheimer's Disease. In this sense, it is important to discover patterns of evolution that distinguish MCI that is stable from those with Alzheimer's in the future. When it is possible to know the trajectories of disease, it also is possible to act more accurately and precisely for each patient. This thesis's main goal is to model the progression patterns of Alzheimer's disease through neuropsychological

tests using of temporal information in earlier prognosis. The main contributions are:

1. Prognosis of AD using a single time point
2. Prognosis of AD using multiple time points (longitudinal information), by exploring time-series summarization techniques
3. Prognosis of AD using longitudinal information, by relying on time-series representation techniques

For this, we initiate the work by reviewing which works are already done. Also, investigate which approach has been using and developing in this matter.

1.2 Thesis outline

This thesis is organized as follows. In chapter 2 describes the background of Alzheimer's disease, Mild Cognitive Impairment, and neuropsychological tests. Next, we introduce the basic concepts of data mining algorithms and techniques. This chapter ends with a related work section where are presented a literature review of AD prognosis problem with machine learning techniques. Chapter 3 describes the dataset used in this work, Complaints Cognitive Cohort (CCC), and describes the time-window approach applied in this work. Chapter 4 describes all preprocessing methods used in this work and the frameworks constructed and followed. This chapter describes the static approach and Temporal approach like summarization techniques, mean and median, and representation techniques, such as Discrete Wavelets Transform (DWT), Signatures methods (ESiG), Symbolic Aggregate Approximation (SAX), and SAX - Vector Space Model (SAX-VSM). Chapter 5 presents all results and discusses this work, namely prediction with static data and prediction with temporal data. Finally, the conclusions are presented in Chapter 6, where some possibilities regarding future work are also presented.

Chapter 2

Background and Related Work

2.1 Overview of machine learning techniques

2.1.1 Missing value imputation

The problem of missing values is transversal to databases, with particular relevance in databases with real values, and this problem can have different fonts. The imputation of missing values is one of the most critical steps in the data processing. This stage's primary goal is to reduce missing values in data, identify outliers, and correct the data's inconsistencies. There are several techniques to handle these problems[21], such as:

1. Do nothing: this method leaves for the differents algorithms to handle the missing value.

2. Imputation using Mean/Median values: this method can only be used by numerical values, replacing the missing value within media/median of the non-missing value in a feature, and performing this independently from other features.

3. Imputation using most frequent or zero constant values is the most frequently useful technique for categorical value. The missing values are replaced with zeros or any other constant value.

4. Imputation using K-Nearst Neighboor: The algorithm uses the 'features similarity' to predict any new data point's values. The missing values are assigned a value based on the closest example, or k-examples, in the training set.

2.1.2 Class Imbalance

The class imbalance is when the data have significant differences between the number of instances of classes. The main problem of class imbalance is that the supervised learning methods assume that the

class labels have the same distribution in data and produce a model biased towards the over-represented-data. The less-represented class goes unlearned. We can apply different techniques that resampling the classes by oversampling the minoraty class or undersampling the majority class to provide this problem. An approach that is very used for this is Synthetic Minority Over-Sample TEchnique (SMOTE). This approach generates more samples from the minority class by choosing similar instances and perturbing the attributes by a random amount[8].

2.1.3 Feature Selection

The time series many times are extensive and have many redundant features or without useful meaning. The main goal of feature selection is to reduce the size of time series, remove the redundant features, and diminish noise from data. This technique results in more critical information and more interpretable values and prevents data overfitting[48].

Feature selection has two main categories: filtering and wrapper methods.[20] The main differences between them are the presence or absence of an algorithm in the selection process. The filter methods choose the feature based on data characteristics (for example, the feature correlation). On the other hand, the Wrapper methods choose the feature based on the importance of classifier performance. The output of these techniques can be a subset of features, returns a subset of original feature cohort, or the original features cohort ranked by their worthiness.

2.1.4 Classifiers

The classification is an important data mining task and has one of the most relevant applications: fraud detection, target marketing, performance prediction, medical diagnosis, and so on. The classification approach has two main phases, the learning phase and the prediction phase. In the first one, we used data to build a model. The second step is to evaluate models' performance, making a prediction of the class label for given data. The following points describe some example of algorithms, such as Naive-Bayes (NB), Decision Tree (DT), Support Vector Machines (SVM), and Neural Networks (NN).

Naive-Bayes

Naive-Bayes are a statistical classifier that made a prediction based on the probability of a given object belonging to a class, using the Bayes theorem represented in the equation (2.1).

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} \tag{2.1}$$

This classifier assumes that the attribute value on a given class is independent of the other attributes, called class-conditional independence. In Bayesian terms, H is some hypothesis, such as object X belonging to a class C_i, and X is evidence, our instance with the attributes. In this sense, the classification problem we want to know P(H|X), the probability of X belongs to a class C_i, given attributes description

of X. This probability is a posteriori probability, H conditioned on X. To achieve this probability, we have to know the apriori probability of H, P(H), the probability of class C_i independent of X attributes. Similarly, we have to know the apriori probability of X, P(X) is the probability of an instance with those attributes in our universe of data. To calculate P(H|X) is necessary P(H), P(X), and P(X|H), which are probabilities that can be estimated from data.

Assume D is a training set of objects and respective class labels, where each object is represented by n-dimensional attributes vectors

$$X = (X_1, X_2, \cdots, X_n). \tag{2.2}$$

Suppose that there are m classes C_1, C_2, ... , C_m. Given X, the classifier will predict which class X belongs, the class with the highest posteriori probability, conditioned on X

$$P(C_i|X) \leq P(C_j|X), for 1 \leqslant j \leqslant m, i \neq j. \tag{2.3}$$

The class C_i for which maximized of P(C_i|X) is called the maximum posteriori hypothesis, by the theorem (2.1).

P(X) is constant for all classes; only P(X|C_i)P(C_i) needs to be maximized.

To compute P(X|C_i), the naive assumption of class-conditional independence is made

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i) = P(X_1|C_i) \times P(X_2|C_i) \times \cdots \times P(X_n|C_i). \tag{2.4}$$

Probabilities quickly retired from training tuples. In the case of categorical attributes, P(X_k|C_i) is the number of objects of class C_i in D having that value for X_k, divided by |C_i,D|, the number of tuples of class C_i in D. On the other hand, if there is a continued value, we assume Gaussian distribution and assume, with mean μ and standard deviation σ

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.5}$$

So its equal to

$$P(X_k|C_i) = g(x^k, \mu_{C_i}, \sigma_{C_i}). \tag{2.6}$$

The predicted label is the class C_i, for which P(X|C_i)P(C_i) is maximum, calculate which means

$$P(X|Ci) \times P(Ci) > P(X|Cj) \times P(Cj), for 1 \leqslant j \leqslant m, j \neq i. \tag{2.7}$$

To predict the class label of a test instance the equation (2.4) is applied for each class, and then is chosen the maximum posteriori hypothesis.

Decision-Tree

The decision tree (DT) class of algorithms are very used in classification problems because they are easy to construct and interpret. These algorithms are a flowchart-like tree structure, where each internal node (non-leaf node), represented in figure 2.1 by circles, denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node), represented in figure 2.1 by rectangles, holds a class label [21]. When we have a new instance without a class label, the classifier tests instance in the decision tree, attribute by attribute, until the leaf node with the class. In figure 2.1, the first attribute is split in two, when we got the answer1 the decision tree continues to attribute 2, and when we have the answer2 we have a class label for that instance. This process is repeated until we have a class label for the instances.

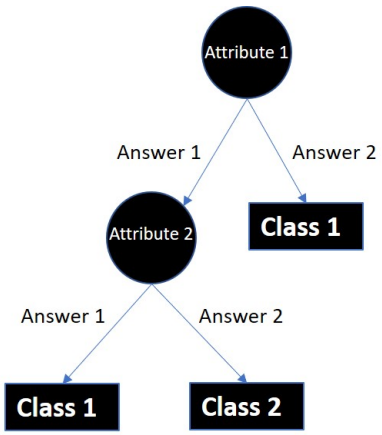

Figure 2.1: Representation of Decision Tree algorithm.

Many algorithms exists to built the DT, such as ID3, C4.5, CART, and SPRINT. The ID3 algorithm devolved by Quinlan, uses the information theory to construct the tree and use the entropy to choose the feature to split. The C4.5 is an improvement of the previous algorithm that can deal with noncategorical data. The SPRINT algorithm can deal with a large amount of data. An attribute selection measure is a heuristic for selecting the splitting criterion that 'best' separate a given dataset. Ideally, the 'best' splitting criterion is that put all the tuples that fall into a given partition would belong to the same class. Then we describe the most favored attribute selection measure: Information Gain, Gini Index, and Gain Ratio[21].

Information Gain

This measure is based on information theory and chose the attribute with the highest information gain (2.8). The attribute chosen minimizes the information needed to classify the tuple in the next partitions. The method reduces the amount of partition and guarantees that the simple tree is found. The information

needed to classify a tuple in D, known as entropy of D, is given by:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i), \tag{2.8}$$

where p_i is the nonzero probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_i|/|D|$. Now we need to know how much information is needed to finalize the classification. For this, we need to calculate $Info_A(D)$ given by:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D) \tag{2.9}$$

A smaller value to $Info_A(D)$ is the attribute that gives us the greater partition. The information Gain, tell us how much we gained by use that attribute for partition, and is defined by differences between original information required and the new requirements after partition, given by:

$$Gain(A) = Info(D) - Info_A(D) \tag{2.10}$$

.

Gain Ratio

The Gain Ratio is a measure preferred to select attributes with many values because it reduces biased toward tests with many outcomes. This method is a successor of information gain. The information gain ratio is calculated by:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}, \tag{2.11}$$

where the SplitInfo(A) is the potencial information generated by splitting the training dataset.

Gini Index

The Gini index method measures the impurity of a data partition or set of training tuple and is given by:

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2, \tag{2.12}$$

where p_i is the probability of an example of D belonging to a class C_i. This technique consideres each of the possible binary split for each attribute, in the case of discrete-valued. The sum of impurity of this split is calculated for all possibles binary split by

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{2.13}$$

With continuous-valued attributes, each possible split may be considered, and the midpoint between each pair is taken as a possible split-point. The attribute split with a more significant reduction in impurity is selected as the splitting attribute. This reduction in classification occurring by applied a binary split on an attribute is given by:

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \tag{2.14}$$

Support Vector Machines

The support vector machine algorithm (SVM) is a classifier that find a decision boundary whose margin is as large as possible, as shown in figure 2.2.

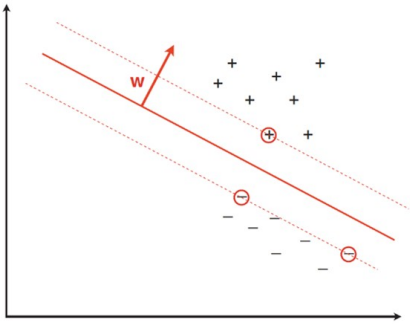

Figure 2.2: Represention of a Support Vector Machine example, and respective decision boundary, and the support vectors.

The SVM finds a decision boundary that separates data from two classes using the essential training tuples and the margins, defined by support vector, represented in figure 2.2 by the dashed line. The margins minimize the classification errors for previously unseen instances and guarantee the largest separation between classes. This separating hyperplane is given by:

$$W \times X + b = 0 \tag{2.15}$$

Where W is a weight vector and b is a scalar. The hyper-plane with maximal margin divide instance into positive, above hyper-plane H1, and negative, bellow hyper-plane H2, and those can be defined by:

$$H1 : W \times X + b \geqslant 1 \tag{2.16}$$

$$H2 : W \times X + b \leqslant -1 \tag{2.17}$$

To construct the margins is used the distance between the decision boundary and nearest training instance from the different classes, called support vectors. In figure 2.2 the support vector was represented with circle points.

Neural-Networks

Neural networks were created to behave computationally analogs to neurons. In [21] neural networks are described as a set of connected input/output units. The multilayer feed-forward neural networks consist of inputs-layer, one or more hidden layers, and an output layer, as represented in the figure2.3.

The units are different in these layers. In the input-layer are units, in the hidden-layers are neurondes, and in the output-layer are output units. The user defines the network topology before training began.

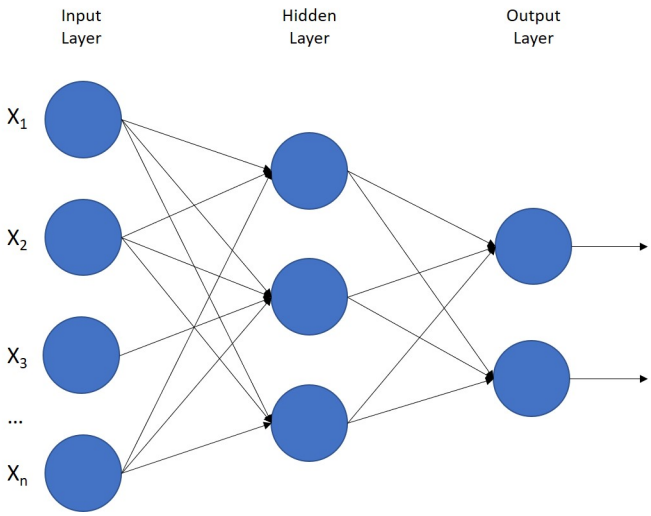

Figure 2.3: Neural networks representation. Example of neural network, this is composed by: one input layer, one or more hidden layers and an output layer

A neural network user has to establish the number of units in the input layer, the number of hidden-layers, one or more, the number of units in each hidden layer, and the number of units on the output-layer. Additionally, algorithms initiate with little random values for weight and bias to these values established by the user. The first step, the learning phase, is the networks making adjustments in weights and predicting the correct class label of the input tuples. Initially, the data pass by the input layer and are weighted and fed simultaneously to the second layer of "neuronlike" units, knows as the hidden layer. Given a neurode j in a hidden or output layer, the net input, I_j to the neurode is

$$I_j = \sum_j W_{ij} O_i + \theta_j \tag{2.18}$$

Where W_{ij} is the weight of the connection from neurode i to neurode j; O_i is the output of the neurode i from the previous layer; and θ_i is the bias of the neurode, which allow for varying the neurode activity. Each neurode in the hidden layer or output layers uses an activation function. The hidden layer's output can be an input of another hidden layer or be an input of the output layer, which gave us a network prediction for a given tuple. The backpropagation is the neural network algorithm more commonly used. This algorithm learns by iteratively processing a data set of training tuples, comparing algorithm prediction with the real known target value. For the classification problem, the target value is a know class label. To a numeric prediction, the target value is continuous. For each training tuple, the weight is adjusted to minimize the mean square error (MSE), defined by:

$$Err_j = (O_i - T_j)^2 \tag{2.19}$$

The weight adjustment minimizes MSE between the predicted target O_i and the real target value T_j.

The algorithm stops when weight eventually converges. This algorithm's principal advantages are high tolerance of noisy data, ability to classify patterns on which they have not been trained, may be used when there is little knowledge of the relationships between attributes and classes. When compared with DT has an advantage in the use of continuous-valued inputs and outputs. A significant disadvantage is the poor interpretability of the process during the training phase.

2.1.5 Evaluation metrics

Several metrics exists to evaluate a classifier's performance and defined how good or how 'accurate' the classifier is at predicting labels of tuples. To define some of these metrics is necessary to clarify some concepts, such as:

- True Positive (TP) - sample classified correctly to a positive class
- True Negative (TN) - sample classified correctly to a negative class
- False Positive (FP) - sample classified incorrectly to a positive class
- False Negative (FN) - sample classified incorrectly to a negative class

With these concepts, we can define some measures like accuracy (or recognition rate), sensitivy (or recall) and specificity.

Accuracy defined by

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.20}$$

is the portion of all correctly positive and negative case in all cases. This measure should be defined with the test set, the class-labeled tuples that were not used to train the model, to avoid over-optimistic estimates.

The sensitivity and specificity measures are highly used to distinguish when the classifier is suitable for detecting positive and negative classes.

Sensitivity, or true positive rate, is the proportion of positive cases correctly classified in all positive cases and is defined by:

$$Sensitivity = \frac{TP}{TP+FN}. \tag{2.21}$$

Specificity, or true negative rate, is the proportion of negative cases correctly classified in all negative cases and is defined by:

$$Specificity = \frac{TN}{TN+FP}. \tag{2.22}$$

2.2 Summarization and Time Series Representation

2.2.1 Time Series Summarization

The summarization method uses the principal characteristics of a time-series to transform it into a new dataset. The principle advantage is significant dimensionality reduction. There is some summarization techniques like statistics-based, Fractal Dimension-Based, and Run-Length-Bases Signature.

The statistic based summarization like mean technique shows the average value of the time-series value and is defined by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{2.23}$$

The statistics median technique shows the values in the middle when putting in an order set of time series values if the number of instances considered is odd, and the mean of two values in the center when the numbers of values are even.

2.2.2 Time Series Representation

Time series are essentially high-dimensional data. It can be complex to find algorithms that work directly on the raw time series. The main goal of a time series representation is thus to emphasize the essential characteristics of the data. The benefits of this strategy are efficient storage, speedup of processing, and implicit noise removal. These basic properties lead to the following requirements for any representation like said in [36]:

1. Significant reduction of the data dimensionality
2. Emphasis on fundamental shape characteristics on both local and global scale
3. Low computational cost for computing the representation
4. reconstruction quality from the reduced representation
5. And insensitivity to noise or implicit noise handling.

Already exists numerous representation techniques. It is possible to divide into three categories according to the kind of transformations applied. The three representation categories are non-data representation adaptative methods, data-adaptive representation methods, and model-based methods. In this work, we only use non-data representation adaptative methods and data-adaptive representation methods.

In Non-adaptive representation methods for representation, the transformation parameters remain the same for every time series regardless of its nature. This work will use only one non-adaptive representation method, Discrete Wavelet Transform (DWT).

Discrete Fourier Transform

The Discrete Fourier transform (DFT) is a representation technique that represents a time series in the frequency domain. The DFT coefficients F_k of a time series $X = \{X_0, X_1, \ldots, X_{n-1}\}$ are complex numbers given by:

$$F_k = \sum_{i=0}^{N} X_i e^{-\frac{j2\pi ik}{N}} \tag{2.24}$$

The adavantages of this technique are:

1. A fast algorithm exist for its computation, called fast fourier transform.
2. Merely the first few coefficients, which correspond to the lower frequencies in the time series, need to be kept for an adequate representation of most time series.

For example, as in [Vla05], using the DFT on a random walk data-keeping only 10% of the coefficients results in retaining of 90% of the energy.

Discrete Wavelet Transform

In wavelet analysis, the Discrete Wavelet Transform (DWT) decomposes a signal into a set of commonly orthogonal wavelet basis functions. Instead of using a fixed set of basis functions, the DWT used dilatated, translated, and scale versions of a mother wavelet function [Chan and Fu 1999], which allow localization of a time series in both frequency and space. DWT is invertible, as DFT so that the original signal can be recovered from its DWT representation. The method allows the analysis of a time series at different scales, as known as resolutions. The DWT use a real-valued vector X of length 2^n, and results in a transformed vector W of equal length. The first step is filtered with some discrete-time, low-pass filter (LPF) h of a given length (in the example in figure 2.4 is used a filter of length four) at intervals of two. Figure 2.4 shows a vector with a length of 16, the first filter resulting in values sorted in the first eight elements of W.

In the following step, the vector X is filtered with some discrete-time, high-pass filter (HPF) g of a given length (again in the example of figure 2.4(b) is used a filter of length four) at intervals of two. This second filter resulting high-pass values are sorted in the last eight elements of W.

The first filter is essentially a down-sample version, and the high-pass part detects and localizes high frequencies in X. This representation gives us a multi-resolution frequency decomposition and localization of a one-dimensional signal, where low frequencies are measured over larger intervals, thus providing better accuracy [39].

Symbolic Aggregate Approximation

The Symbolic Aggregate Approximation (SAX) implies that a transformation's parameters are modified depending on the data available. This characteristic is a potent advantage relatively to non-data adaptive

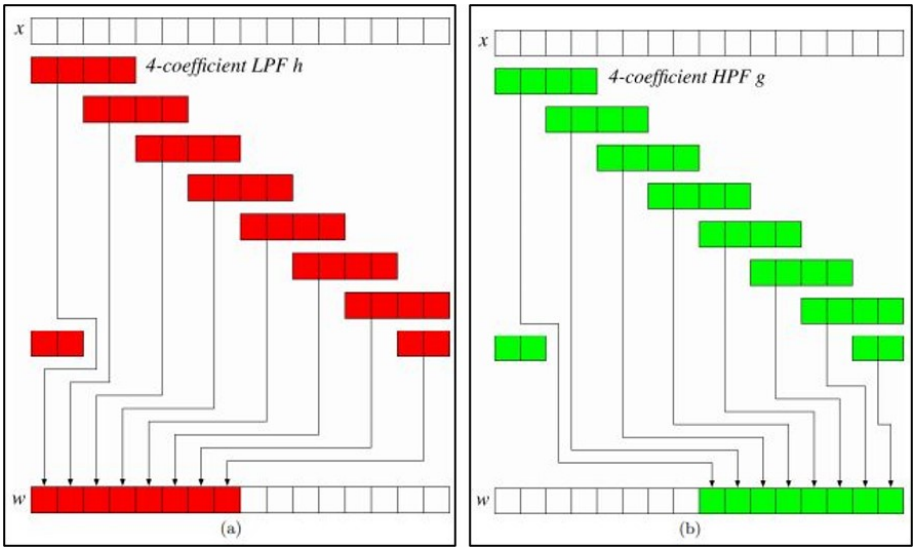

Figure 2.4: Representation of a transformation based on DWT. The image A represents the first transformation with low-pass filter. The imagem B represents the transformation with hih-pass filter.

because this is very powerful with missing value on time series, which are frequent in real data. However, it is a step that can be added to almost all non-data-adaptive method The SAX[28] calls on equal frequency histograms on sliding windows to create a sequence of short words. SAX is a classical symbolic approach for time series data mining. This technique fundamental is to convert the numeric values into discrete symbols to designed mapping rules. This time series representation reduces a time series of arbitrary length n to be a string of arbitrary length w. The alphabet size used is optional, but bigger than 2[28]. In the first step, SAX transforms data into Piecewise Aggregate Approximation (PAA) and then symbolizes the PAA representation into a discrete symbolic string. The dimensionality reduction with PAA reduces the time series from n dimensions to w dimensions. This method divides the data into W equal-sized frames, as shown in figure 2.5 where the time series C transforms into C'. The representation of frames is the average of the values. A new vector with these values becomes the data-reduced representation.

This dimensionality reduction process has been shown to rival more sophisticated dimensionality reduction techniques like Fourier transforms and wavelets [22,23,25], also used in this work. In the next step and having a time series database into PAA form, we can also perform other transformations to obtain a discrete representation. All coefficients obtained in PAA are attributed to a symbol. The symbol commonly used is the first letters of the alphabet, for instamce {a, b, c, d}, that difined the limit of a breakpoint. If this value of coefficients is below, the first breakpoint is assigned to a symbol 'a'. For values between the first and second breakpoint are assigned to the second symbol 'b'. For values between the second and third breakpoints are mapped to 'c'; all values greater than the third breakpoint

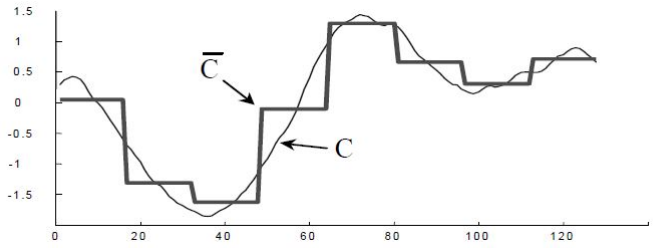

Figure 2.5: Representation of PAA process in reducing features. The PAA representation example of a sequence with length 128 transformed into eight dimensions

are assigned to ‘d’. The figure 2.6 illustrates this representation with only two breakpoints.

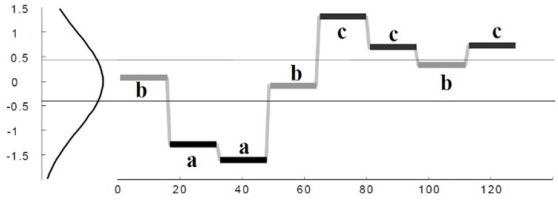

Figure 2.6: Representation of SAX transformation. The SAX transform the real values of a time series into a sequence of letter.

We obtained a new representation of data with a sequence of letters representing the original values in time series. In the example represented in figure 2.6, the sequence obtained is ‘baabccbc’.

SAX-VSM: Symbolic aggregate approximation and Vector Space Model

The Symbolic Aggregate Approximation-Vector Space Model automatically discovers and ranks time series patterns by their importance to the class label. Pavel’s work shows an alternative to the nearest-neighbor algorithm (1NN) with superior interpretability, learns efficiently from a small training set, that are our case, and has a low classification computational complexity. This algorithm looks for time series subsequences, which represent a class that enables superior interpretability. This algorithm ranks by importance all potential candidate subsequences at once with a linear computational complexity of O(nm). SAX-VSM converts all training time series into bags of SAX and uses tf*idf weighting and Cosine similarity. This approach uses two techniques that are very well known, SAX, which is a high-level symbolic representation of time series [43, 29], and VSM that is based on tf*idf weighting scheme:

1. SAX transforms time series real values into combined collections of SAX words
2. Using $tf \times idf$ weighting, it transforms those collection into class-characteristics weight vector
3. Values obtained in point 2 is used in classification built upon Cosine similarity.

Figure 2.7 shows an example of this transformation, with a time series, following the SAX representation, and finally, the calculation of tf*idf to discover a pattern.

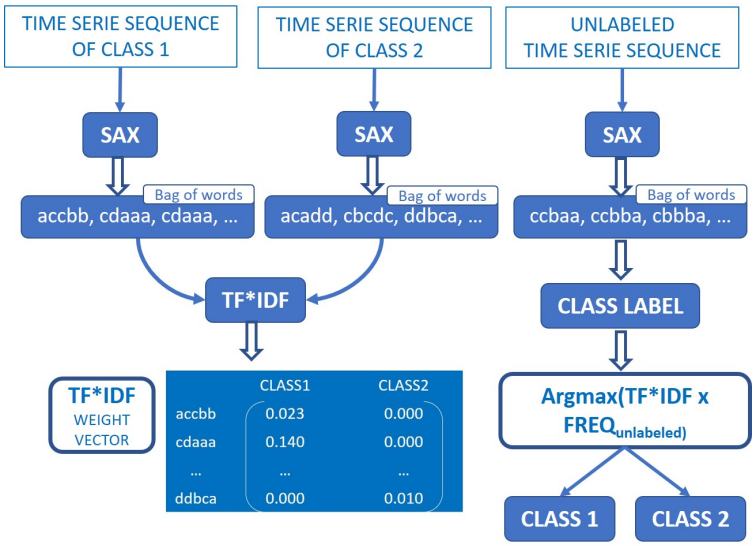

Figure 2.7: The SAX-VSM first transforms data into a sequence of letters. After this step apply TF*IDF to compare all sequence and calculate similarity between sequences.

Signature Transformation

The signature transformation turns sequential data into useful new features. This technique presents high robustness to downsampling, showing identical results in transformation with time-series with 100 points and reduced version of the same time-series with 20 points. Consider a time-series X of finite length in d dimensions which can be described by

$$X = X_t^1, X_t^2, X_t^3, \cdots, X_t^d \tag{2.25}$$

where each value X_t^i is real-valued and parametrized by t∈[a,b]. The signature transformation S of a path X is defined by

$$S(X) = \left(1, S^{(1)}, S^{(2)}, S^{(1,1)}, S^{(1,2)}, S^{(2,1)}, S^{(2,2)}, \cdots, S^{'}\right), \tag{2.26}$$

as an infinitive collection of terms. Which terms of this collection is a k-fold iterated integral of X with multi-index $i_1, \cdots, i_k$, defined by,

$$S^{'} = \int_{0<u_1<u_2<u_k<t} dX_{u_1}^{i_1} dX_{u_2}^{i_2} \cdots dX_{u_k}^{i_k} \tag{2.27}$$

The new features represented by S(X) completely characterize a path X. The usefulness of signature as a feature of sequential data is mostly used in different domains and was demonstrated for many machine learning applications such as healthcare[24, 23], finance[4], data analysis[9], and deep signature learning[37].

To better understand this method, we use the following example to present how intuitive this tool can be and demonstrate the robustness when the timepoints are reduced. This example made a representation of calculus with a short time series. The result of successive integral, calculated by:

$$S^{(1,2)} = \int_0^T \left[\int_0^{T2} dX^1(t_1) \right] dX^2(t_2) \tag{2.28}$$

and

$$S^{(2,1)} = \int_0^T \left[\int_0^{T2} dX^2(t_1) \right] dX^1(t_2) \tag{2.29}$$

is the area's value of $S^{1,2}$ and $S^{2,1}$. The line constructed from short time series points, shown in the figure2.8, and we can see that if the line has more points of the time series the area do not change significantly.

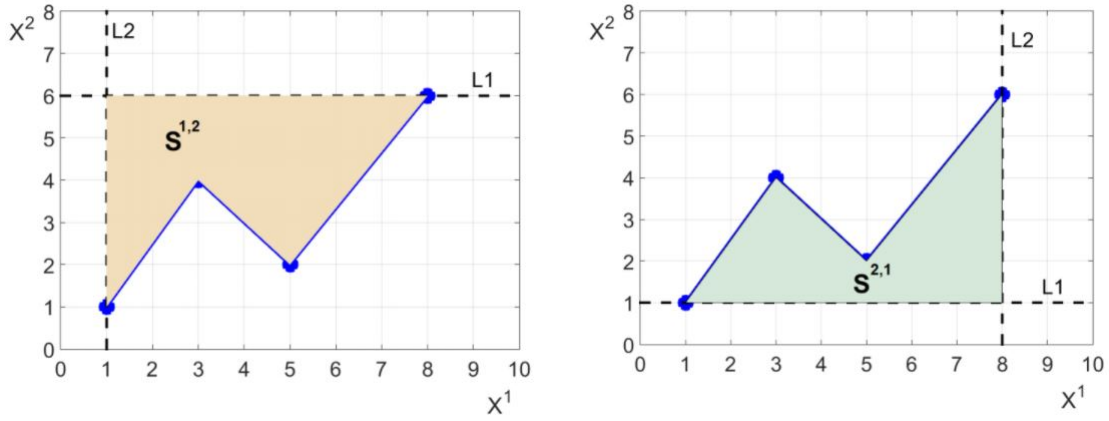

Figure 2.8: Example of signature method

2.3 Neuropsychological Tests

The diagnosis of Alzheimer's (AD) disease is made through neuropsychological tests, Magnetic Resonance Imaging (MRI), Positron Emission tomography, or analysis of Cerebrospinal Fluid (CSF). This work will be using neuropsychological tests. The neuropsychological tests allow collecting information on a patient's symptoms and performing a neuropsychological examination to identify clinical signs and identify a lesion in the nervous system's specific location. These tests have been indicated when detailed information about cognitive function will aid clinical management:

1. To assess the presence or absence of deficits and to delineate their patterns and severity;
2. To help to establish a diagnosis (e.g., Alzheimer's disease or frontotemporal dementia) or to distinguish a neurodegenerative condition from a mood disorder (e.g., depression or anxiety);
3. To clarify the cognitive effects of a known neurological condition (such as, multiple sclerosis, stroke or brain injury).

Neuropsychological tests also can help in a (differential) diagnosis, obtain prognostic information, monitor cognitive decline, control the regression of cognitive-behavioral impairment in reversible disease, guide prescription of medication, measure the treatment response or adverse effects of treatment, define a baseline value to plan cognitive rehabilitation or to provide objective data for the medico-legal situation. ThusThese tests provide general and specific information about cognitive performance. The points below show the multiples examples of neuropsychological tests:

1. Brief state examination provides a quick and easy global, although a rough, measure of a person's cognitive function.
2. Comprehensive neuropsychological evaluation explores several cognitive domains (perception, memory, attention, executive function, language, motor, and visuomotor function).
3. Also, single tests are designed to explore a specific domain or subdomain preferentially, but most of them examine multiple cognitive functions. For this reason, neuropsychological tests are performed as a battery, with more than one test for each cognitive domain.

There are many types of neuropsychological tests for measuring cognitive decline. The most used tests are the Alzheimer's Disease Assessment Scale (ADAS) and the Mini-Mental State Examination (MMSE). Also, many more tests exist, like Montreal Cognitive Assessment (MoCA), Cambridge Cognitive Examination (CAMCOG), Ray Auditory Verbal Learning Test (RAVLT), California Verbal Learning Test(CVLT), among others. The ADAS is a cognitive test designed to measure the severity of most symptoms of AD. This test comprises eleven tasks measuring the problems related to memory, language, praxis, attention, and other cognitive abilities related to Alzheimer's disease symptoms. ADAS-cog measures have been used in different clinical trials and different diseases, such as Mild Cognitive Impairment, Vascular Dementia, and Parkinson's disease. When the total score of this test is low indicates better cognitive performance. The MMSE is a 10 minutes test of flawed thinking in an adult population. This test indicates the severity of cognition impairment. This test's score range varies from 0 to 30, and lower scores correspond to severe cognitive impairment. The domains that this type of test consider are orientation, registration, recall, calculation and attention, name, repetition, comprehension, reading, writing, and drawing. The repetition of this test and decreasing scores can indicate the deterioration in cognitive. In the literature there are different studies that use only one cognitive test alone [26, 12, 47, 31, 46], and in combination with another cognitive test [32, 38, 41, 10]. Still one can find research about combinations of different ways of diagnosis using biomarkers and cognitive tests [17, 18, 1, 25, 11].

These studies focus on analyzing these tools' ability to diagnose and predict neurodegenerative diseases' progression, like a progression from MCI to AD. Also, many studies demonstrate the importance of the neuropsychological test in diagnosing the disease, comparing the different neuropsychological tests for the detection and progression of MCI to Alzheimer's disease [33, 35] and its differentiation from other ones [34]. In [34], a battery of neuropsychological tests was validated to detect MCI and can distinguish AD from frontotemporal dementia (FTD). In the study of Eckersom [13], it compared three diagnoses: independently, neuropsychological tests, Cerebrospinal fluid markers, and neuroimaging and its combinations. The results reveal the importance of neuropsychological tests. Memory test RAVLT was considered the best predictor with an AUC of 0.93. Its combination with the other methods improved the predictive ability (AUC of 0.96). The study of Li et al., [27] use five different neuropsychological tests, ADAS-cog, RAVLT, MMSE, MoCA, and CDR sum of boxes (CDR-SB), to compare the ability to evaluate the progression of the disease with images from MRI exams. They conclude with this study that cognitive measures were significantly stronger predictors than imaging measures, and ADAS-13 has considered the optimal predictor. According to the study by Restaino et al., [41] cut-off scores from CAMCOG memory subscale predicted dementia with acceptable accuracy. In contrast, the non-memory scale scores had low accuracy and consequently did not recommend its use for predicting high-risk cases unless all of the non-memory sub-domain are combined.

2.4 Machine Learning in Alzheimer's Diseases

Several studies use different machine learning (ML) algorithms, such as Naive-Bayes, Decision trees, and Support Vector Machines. Since many studies use these techniques, we also start our work with them. ML algorithms have been used to compare different tests' ability to diagnose or predict disease development or compare with other ones to conclude which algorithms have the best accuracy in these predictions. In recent years different machine learning techniques have been applied to diagnose AD, and these techniques have been applied to different types of data, mostly in data from baseline (data from first assessment). For example, in the study of Esmaeilzadeh et al. [14], they used 3D convolutional neural networks with 3D images and have good results, the accuracy of 94.1%. Long et al. [30] uses a Support vector machine based on magnetic Resonance Imaging (MRI) scans dataset with an accuracy of 96.5%. Also, Zhang et al. [51] use Support Vector Machine, but with more information from MRI scans, PET scans, and CSF biomarkers to diagnose AD with an accuracy of 93.2% in the prediction of AD from individuals controls.

2.4.1 Machine learning in Alzhiemer's Disease with neuropsychological tests

Many studies performed machine learning algorithms with information only from neuropsychological tests [26, 12, 47, 31, 46, 32, 38, 41, 10], similarity to our approach. Some studies use one algorithm to compare different neuropsychological tests, like in [44]. The study by Shankle [44] used the decision

tree algorithm to compare two different neuropsychological tests, Functional Activities Questionnaire (FAQ) and the Six-Item Blessed, Orientation, Memory and Concentration Exam (BOMC); both are recommended for screening by the Agency for Health Care Policy Research (AHCPR). Using C4.5 and C4.5 rules, they obtained 85.5% and 85.9% for the accuracy (probability of overall correct classification).

Moreover, many studies also use more than one ML method with clinical data, purposing to diagnose the disease as earlier as possible [2, 19]. In both of these studies is made an exploration with SVM and Naive-Bayes. In [2], they develop a novel preprocessing approach, and [19] develops an algorithm for a 3-year prediction of conversion to AD.

Most of the comparative studies using neuropsychological tests use a machine learning algorithm for classification tasks and compare their accuracies. Weakley et al. [50] want to reduce the amount of testing to detect cognitive impairment, and for this, they use Naive-Bayes, Decision Tree, and logistic regression.

Also, Shen et al. [45] use Decision Tree and demonstrate in their study that any neuropsychological test alone was sufficient for the diagnosis of AD. This study identifies that Logical memory tests are the most critical indicator of AD diagnosis. The DT's performance was compared between the reduced features set with an accuracy of 73.9% and sensitivity of 85.0%, and the DT based on all neuropsychological tests, with an accuracy of 73.3% and sensitivity of 84.5%.

2.4.2 Machine learning in Alzhiemer's Disease with Temporal Data

In most scientific fields, measurements are performed over time. These observations lead to a collection of organized data called temporal data. The purpose of temporal data mining is to extract all meaningful knowledge from the shape of data. Nowadays, many studies use machine learning for analyzes of temporal data in very different domains, from predicting the progression of diseases to predicting suicide attempts in adolescents [11, 25, 1, 18].

Regarding Alzheimer's disease, many advances are made in this area. Recently, different approaches are developed to lead and process data. It was already developed methods that lead to data from different clinical assessment results, from neuropsychological tests, MRI, and PET. The study-oriented by Zhu [52], use three different cognitive tests, such as MMSE, ADAS-cog, and CDR-SB, to model and predict the entire clinical trajectory. They present a probabilistic, latent disease progression model for capturing dynamics of the underlying pathology. On the other hand, a new machine learning technique was developed by Bhagwat [6], Longitudinal Siamese Neural Network (LSN). This technique uses MMSE, ADAS-cog, and MRI data from two-time points and can effectively combine them and improve predictive performance. However, in this study, temporal data is only used to train the model, and for validation, baseline data was used. When applied to MMSE an accuracy of 90% was obtained and an AUC of 0.968 for binary MMSE task, and 76% of accuracy for ADAS-13 of the ADNI dataset. In replication on the AIBL dataset was obtained an accuracy of 72.4% and 0.883 of AUC for binary MMSE tasks.

2.5 Summary

This chapter, began by describing the principal concepts of machine learning, emphasizing techniques that we will use in our work. Next, the Alzheimer's Disease is described, as well as the tests used to evaluate a patient, particularly the neuropsychological tests, and an overwiew of the related work is presented. These tests are divided into a series of battery tests that assess and monitor the patient's mental health. The neuropsychological tests have many advantages from the biomarkers. For example, they are more widely applied and are a non-invasive test. It also has disadvantages like a subjective analysis by doctors. The following chapter describes the dataset used in this work.

Chapter 3

Dataset

3.1 Cognitive Complaints Cohort

The data used in this work was from Cognitive Complaints Cohorts (CCC), a predictive study guided by Faculdade Medicina of Lisbon. This CCC intends to study the progression of dementia in individuals with cognitive complaints based on extensive neuropsychological evaluation in one of two institutions (Laboratório de Estudo da linguagem, in Hospital Santa Maria, and Memory Clinic, both in Lisbon). The inclusion criteria had to be fulfilled for individuals who were included in CCC. These criteria consist of cognitive complaints and completing the assessment with a neuropsychological battery to evaluate many cognitive domains and validated for the Portuguese population (Bateria de Lisboa para Avaliação das Demências (BLAD) [15]). The exclusion criteria considered for CCC was a diagnosis of dementia (according to DSM-IV [5]) or other problems that may cause cognitive impairment. Some causes of cognitive impairment can be a stroke, brain tumor, significant head trauma, epilepsy, psychiatric disorders, uncontrolled medical illness, sensory deficits, medical treatments that interfere with cognitive function, and alcohol or illicit drug abuse. For participants with Mild Cognitive Impairments (MCI) included in this work had to be fulfilled the criteria of the MCI Working Group of the European Consortium on Alzheimer's disease [40]:

1. Cognitive complaints coming from the patients or their families;
2. Report of decline in cognitive functioning relative to previous abilities during the past year by the patient or informant;
3. Presence of cognitive impairment (1.5 standard deviations below the reference mean) in at least one neuropsychological test;
4. and Absence of major repercussions on daily life activities.

The different cognitive tests that compose the CCC is BLAD, MMSE, Trail Making Test, CVLT, Geriatric Depression Scale (GDS), Subjective Memory Complaints Questionnaire, and Blessed Dementia Scale. The CCC data was updated, and this work will used the version, of October 2018. Our data comprises 99 variables containing clinical, demographic, and neuropsychological information from patients and results from the clinical assessment from five years of follow-up.

Next, we describe how we construct the class label with the Time-Window approach for two different perspectives, static and temporal analysis. A consideration in our work is using the time factor to obtain more useful information and considering patients' temporal progression. We defined a new class label for the instances for this goal. For that, we used a time-window approach at different duration times. We use this technique with a few differences between static analysis and temporal data analysis. A necessary criterion to be part of a new dataset is to initiate the evaluation at baseline with a diagnosis of MCI. All examples that do not satisfy this criterion were eliminated. The next step is to define the time considered by the time-window. This step is different between static analysis and temporal analysis.

3.2 Time-Window approach

For the Static analysis, we chose time-window approaches with two, three, four, and five years. The following figure demonstrates this approach.

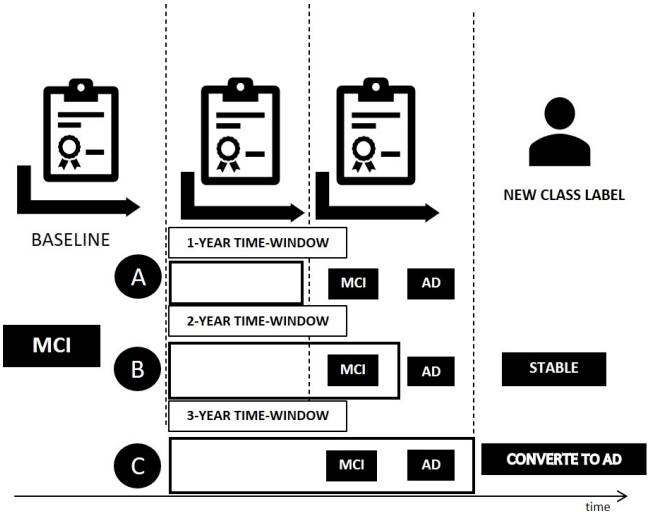

Figure 3.1: Creation of learning examples based on TimeWindow approach [38] with static data.

Figure 3.1 has three possible examples of class labels for the one patient person. This technique assumes that the new class label depends on the last medical appointment's diagnosis results inside of time-window. For instance, to be considered a converted example (cMCI or AD), the patient needs to

be diagnosed as AD inside the time-window period in consideration. The following instance represents the successive diagnosis along the time. In this example, the one-year time window does not have a new diagnosis for that instance represented by A in figure 3.1. Example A in figure 3.1 can happen when two medical appointments are separated by more than a year. In this case, this person will not be considered with in a one-year time-window. When we use a two-year time window approach, represented in example B in figure 3.1, we can see that the confirmed MCI diagnosis is obtained inside the period considered by this time window. So, in this case, the person will be considered stable MCI. The last example of figure 3.1 represented by C, is a three-year Time-Window. In this case, the last diagnosis inside the box is the conversion to AD. So, the person is considered with the class label of converted to AD in a three-year time-window.

Then we defined the class label for all instances with this approach and used two more time-windows, with four and five years. For those time-windows, the rationale is the same.

For temporal data, the time-window approach is performing after the third medical appointment. Figure 3.2 represents the approach used in a temporal data perspective.

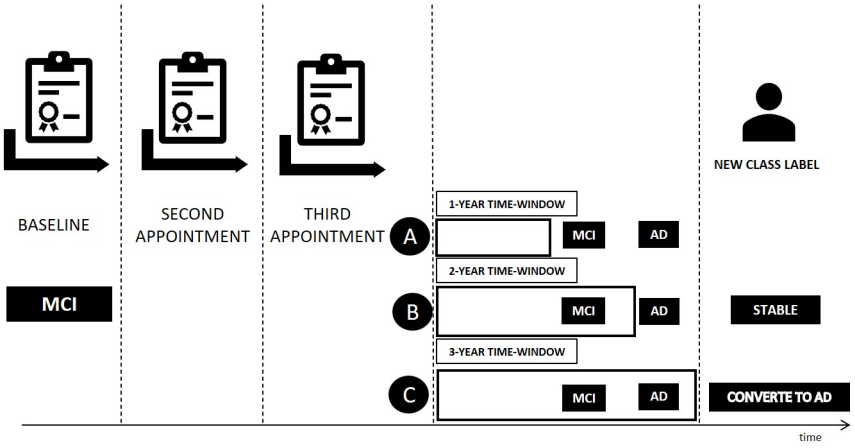

Figure 3.2: Creation of learning examples based on TimeWindow approach [38] for temporal data.

The method to associate the class label is similar to the static data approach. In this case, the dataset is composed of three instants of time. The first appointment is described as the baseline in the figure, the second and third appointments for temporal features, and the static data, the features are stable along the time.

3.2.1 Dataset description

After applying four different Time-Windows for each patient, four different datasets are obtained corresponding to two, three, four, and five years of time-window.

Initially, we performed a statistical analysis of data. Our datasets' demographics descriptions are represented in the table. 3.1.

Table 3.1: The demographic description of static data

Time Window	2-years	3-years	4-years	5-years
Examples	501	468	431	410
Class (0/1)	394 (21.4%) 107 (78.6%)	305 (67.2%) 163 (34.8%)	227 (52.7%) 204 (47.3%)	235 (57.3%) 174 (42.7%)
Gender (1/2)	195/306	182/286	161/270	75/100
Missing value	13933 (28.1%)	12968 (27.99%)	11971 (28.06%)	11636 (28.66%)

The mean age at the beginning is 68.67 $\pm$ 8.78. Also being possible to verify that the gender distribution was not equal in the four datasets, feminine gender appears more than masculine in all of the datasets. It is possible to verify that the example numbers decrease with the increase of time in the time window, with a tendency to stabilize, which is considered normal because fewer people complete a higher number of medical appointments. The number of missing values is always around 28%, which is a high number and is expected when referring to a real dataset.

For temporal data are applied three different time-windows, with one, two, or three years after the third appointment. The main set is equal in three datasets with different class labels depending on the time of time-window. The new dataset obtained has a new statistical analysis, as shown in the table below 3.2.

Table 3.2: The demographic description on temporal data when time-window approach is applied with one, two and three-year.

Time-Window	1-year	2-year	3-year
Examples	109		
Class (0/1)	99/10	91/18	86/23
Gender (1/2)	43 (39%) / 66 (61%)		
Missing Values	15784 (39%)		

The new dataset has a more balanced distribution in gender, and the mean age is 69.86$\pm$ 7.90. These differences are expected because we only use more complete examples, with more than four medical appointments and a more extended medical history. This selection is necessary for taking into account a temporal evolution. The class imbalance is much more evident in this case, with the converted class underrepresented in three time-windows.

Based on the new class balance results, we intend to predict the disease's evolution within three years' time-window. The class label imbalance resulted from this approach is very evident. So, only the

three-year time-window can be used without too many fictitious instances created by SMOTE. For this reason, we only used a three-year time-window. The Time-Window approach has the advantage that it is not static in time. It can be applied with different duration of time-windows and at different times of time-series. So for future works, this approach can be flexible and can be applied in other contexts.

3.3 Summary

This chapter has described the dataset used in this work. The dataset was from Cognitive Complaints Cohort (CCC), a predictive study guided by Faculdade Medicina of Lisbon. Additionally, it is described the inclusion and exclusion criteria for this dataset. Furthermore, it describes the Time-Window approach and the differences when applied to static data and to temporal data. In the next chapter, it will be described the methodology followed in this work.

Chapter 4

Methodology

In this chapter, the methodologies used in this work are described. Figures 4.1 and 4.2 represent the sequence of our approach to handle the original dataset from CCC. This work follows a methodology represented in figure 4.1 as the initial approach regarding the static data. Initially, it was performed static data analysis, only using one time-point and without consider the following consultation evaluation.

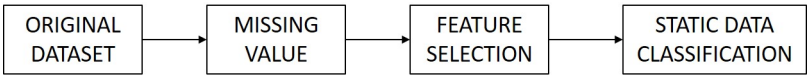

Figure 4.1: Approach used in analysis of static data.

On the other hand, it was performed an analysis with temporal data, which used different approaches, and used techniques that consider time, such as summarization and representation techniques. Figure 4.2 represents the sequence of techniques used in this case.

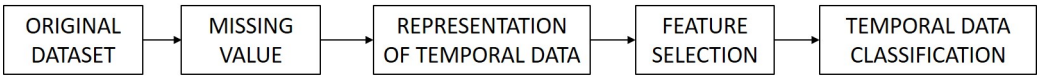

Figure 4.2: Methodology used in temporal data analysis.

The temporal data analysis has one more step between the missing value step and the feature selection step. In the representation of temporal data, it is used techniques of summarization and representation, which results in the new dataset.

The main difference between the two approaches is the representation of temporal data as described later in this chapter before the feature selection steps, to consider three-time points and evolution along the time. For our analysis, the news datasets were obtained as described in section 4.1, to consider only one-time point and consider the factor time in the approach to temporal data.

4.1 Temporal Dataset pre-processing

In temporal data, the initial step is to prepare our dataset to extract the maximal useful information. The summarization and representation techniques initially use the same step to lead with missing values, namely last o-bservation carried forward (LOCF). The last observation carried forward basics means that unless we have a different value to the next observation, we assume the last observation's value. The summarization techniques help us construct a new dataset containing useful information from the original dataset. To obtain a new dataset from the original one, it uses the firsts three time-points from our short time-series. Both summarization techniques used in this work, mean and median, follow the same rules to construct the new dataset. Figure 4.3 explains how to construct the dataset without missing values, or in case that one missing value is in the middle or last position.

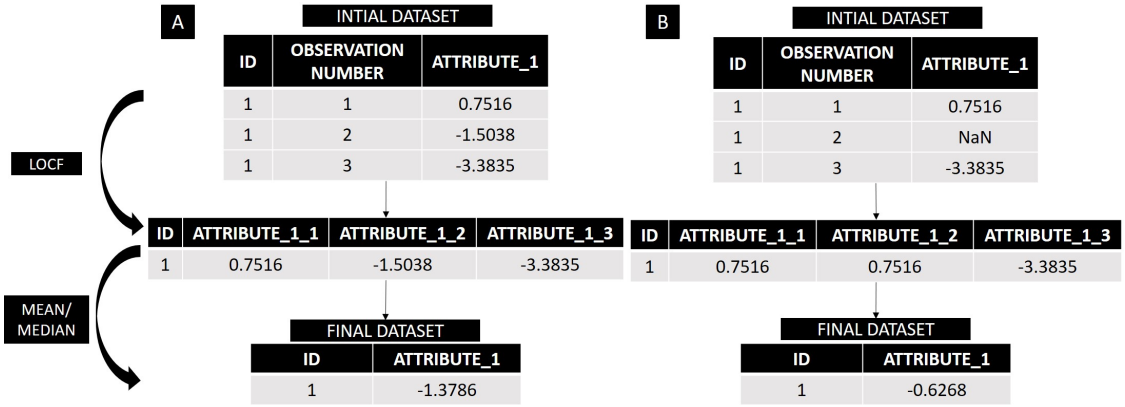

Figure 4.3: Example to obtain a new dataset using summarization methods.

Example A of figure 4.3 represents the first situation where the values for the three observations are known. In this case, three lines in the original data set are transformed into one row for the same person, and the final values are the mean/median of these values. In case of a missing value, excluding missing values in the first medical consultation, it is assumed that there is no information to prove that a change occurred since the last observation. So when a missing value is in the second or third observation, it is assumed the previous observation's value. The example B of figure 4.3 represents these situations. So the three rows of the original dataset transform into one row, and the final value is a mean/median of these three values.

Another situation that can happen is the presence of two missing values. In this case, it is assumed that if we do not have information to the contrary, the value remains the same as the last observation, LOCF. This last situation is represented in the figure 4.4 with example C. On the other hand, the two missing values can be the firsts two observations. In this case, in the absence of value since the first appointment, the missing values take the value of zero. The example D in the figure 4.4 below represents this situation. In both these situations, the final value is the mean/median of values considered after

transformation.

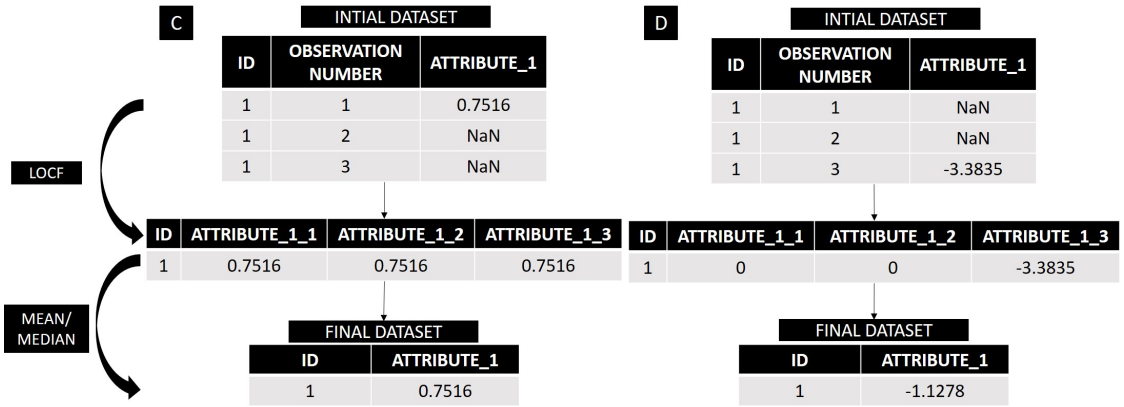

Figure 4.4: Construction of new dataset with summarization technique dealing with missing values.

To obtain a temporal dataset for representation techniques is similar to the dataset obtained with summarization techniques. The main difference is that only the initial transformation from three rows of the original dataset to one row will be done as shown in figure 4.5.

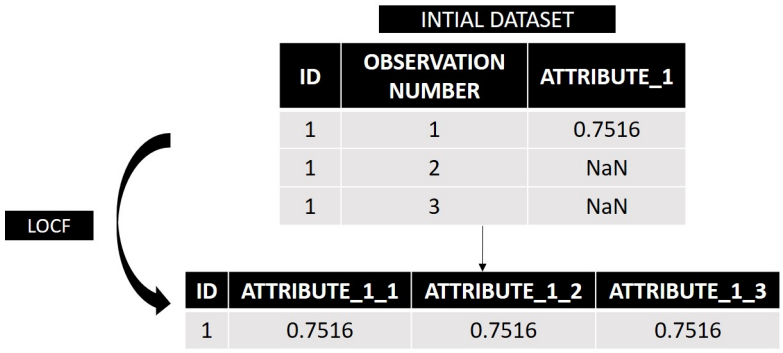

Figure 4.5: Construction of new datasets to be used by representation technique. This figure represents the approach to obtain a new dataset in the presence of missing value.

All the examples of missing values in the representation dataset construction are treated in the same way as described above.

4.2 Prediction on Static Data

The first step in any data analysis is to improve the quality of data and the information that it is possible to extract. Figure 4.6 has the diagram representing all the processes applied throughout this work.

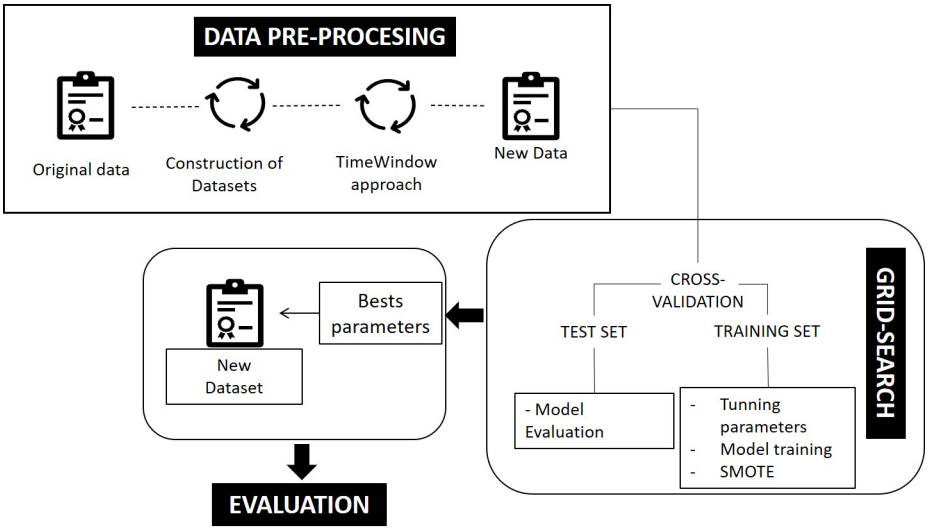

Figure 4.6: Schematic of the proposed methodology.

First, some data pre-processing is applied such as handle with missing values and class imbalance, and perform feature selection. From the original dataset, only values about the first medical appointment was selected when the static data is considered. After the data selection, we have to deal with missing values. In the static data analysis, all missing values of numerical attributes were replaced by the mean, and the mode replaced all missing values of categorical attributes.

The next step is to reduce the number of features and remove the less informative features. The features selection step results in similar features to the work of [38]. To perform FS has used variance threshold techniques and multiple ranking methods. Telma's approach in [38] uses multiple ranking of features, reducing the effect of dimensionality, increasing the discriminating power, and reducing the effect of missing values on results.

The first part of this work used four classifiers, namely Naïve-Bayes, Decision Tree, Support Vector Machines, and Neural networks. All implementations of these classifiers are made with Scikit-learn. One reality of our data is the class imbalance. The training set method to deal with this problem is SMOTE, which can simulate instances of class underrepresented. Before we applied any classifier, except the Naive-Bayes, the best parameters are determined using 10-fold cross-validation with a grid search approach. In this work, the different grid searches are done with the different parameters showed in the table 4.1.

After the best parameters are defined, the performance of the four classifiers is achieved.

This first analysis in static data is used to chose which classifier is better to deal with our dataset. In the second approach, we do a new analysis using temporal data, and only the classifier with best performance and most promising is used, which is SVM.

Table 4.1: Grid search parameters

Classifer	Parameters
DT	Criterion: {'Gini', 'Entropy'} Splitter: {'best', 'random'}
SVM	linear: {C: 0.001, 0.01, 0.1, 1, 10, 100} rbf: {C: 0.001, 0.01, 0.1, 1, 10, 100; gamma: 0.001, 0.01, 0.1, 1, 10, 100}
Neural-networks	Hidden-layer: {1, 5, 10, 15, 20, 25, 30}; Solver: {'lbfgs', 'sgd', 'adam'}

4.3 Prediction on Temporal data

In our analysis with temporal data, we have to change some steps from static data. The main difference between the two analyses is the construction of the dataset. As described earlier, we used three-time points of each patient for each attribute. Another big difference is the time-window approach, also described above. The missing value situation, in this case, is considered from a different perspective. In the case of temporal data, we can not assume that the other individuals' mean can replace one missing value because this can change the value sequence's trend to be increasing or the opposite situation to be decreasing. So we used the last observation carried forward. With this process, we do not have missing value in short time-series, namely mean, median, DWT, ESiG, and SAX. In this step, we construct five different datasets with different summarization techniques and representation of time-series, as described earlier. Next, we applied a time-window approach that attributes the class label for all instances based on a time-window after the third medical observation. In this case, only three years time-window approach is used because SMOTE's application in one and two years time window will produce too many fictitious examples for the undersampling class. The Grid Search step is performed in the same way as in static data but only use SVM after showing the best results in the initial analysis. This time, new evaluation performance is done with the two summarization techniques and the three different representation techniques, with the best parameters.

4.4 summary

In this chapter, we have a description of all methodologies used in this work. Initially, the sequence approach was used in further analysis, static and temporal data. Afterwords we have done a description of the new dataset's construction and an explanation of the construction of class labels using the time-window approach. In the static data, we use a one, two, and three-years of time window. The temporal data with one and two years generates too much imbalance class in data to be used in our analysis. We

only use the three-year time-window approach. In the next chapter, we present all results of our analysis.

Chapter 5

Results and Discussion

5.1 Initial results with static data

For the first analysis, we chose four classifiers and applied them to four datasets. These datasets differ in the construction of the class label. We use the time-window approach described in section 3.2 to obtain the class labels at different times from two years to five years. The first experience is to choose the classifier that better handles our data for a posterior comparison with the same classifier applied to temporal data. The preliminary results are shown in figure 5.1.

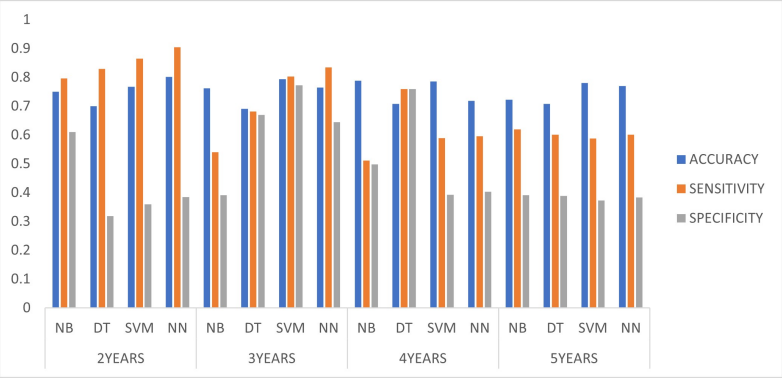

Figure 5.1: The results obtained with the static data classification, in four classifiers and four different time-windows. The classifiers used are Naive-Bayes (NB), Decision Tree (DT), Support Vector Machines (SVM), and Neural Networks (NN).

Figure 5.1 shows that none classifier is significantly better. Almost all classifier has an accuracy between 0.7 and 0.8. Most of them show lower results in sensitivity and specificity. In the second step, we intend to discover the best classifier to dead with our dataset for future discoveries and achieve our goals. For that, we perform the first classification and use a unique time- point. So, we decide to choose

the dataset more balanced and then choose the classifier with better performance. In four years and five years time-window, the differences in the balance of data were not significant. Regarding missing values, the four-year time-window dataset has less missing value and more data than the five-year time-window dataset. For this reason, we used four-year of follow-up dataset for the classification analysis of possible prognostic prediction of AD, an identical approach used in the study developed by Pereira[38]. The classifiers used are Naive-Bayes (NB), Decision Trees (DT), Support Vector Machine (SVM), and Neural Networks (NN). We choose these classifiers because they are the most used algorithms in this type of analysis. In this experience, we obtained the results shown in figure 5.2.

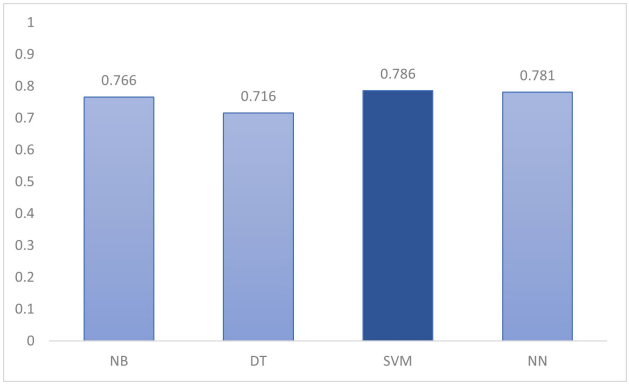

Figure 5.2: Results of classification with static data methods in four-years time-window dataset.

According to figure 5.2, support vector machine is the most promising method to continue the work. Also, we opted to choose only the SVM for the next steps because this is the most used in temporal data analysis, identical to our approach.

At this point, the results of the previous classification experience only consider the data from one time appointment, do not take into account the differences, improvement or worsening, observed in two different follow-up appointments. The main question that we want to answer is if we use information from two different times of follow-up, can we get more specific information and more accurate predictions? In the next step of this work, we will use a different approach with temporal data, which allows us to ensure that the prediction results will take into account more than one moment of clinical evaluation.

5.2 Summarization techniques

The two techniques of summarization that we used in this work are mean and median. Remember that the dataset corresponds to a three-years time-window, which means the method will try to predict if a subject will convert or not in three years. In figure 5.3, we can observe the results with the SVM applied to the test set, with the best parameter obtained in the grid search. In both cases, we have good values of accuracy, with a high value in the median.

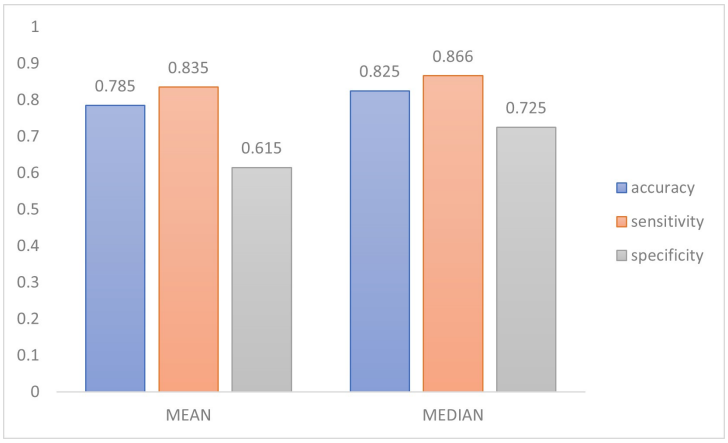

Figure 5.3: Results of classification with summarization methods.

The mean method shows an accuracy of 0.785, a specificity of 0.615, and a sensitivity of 0.835. The median approach produce slightly better results with an accuracy of 0.825, a specificity of 0.725, and a sensitivity of 0.866. The results with the mean dataset are slightly worse than the median results. In both cases, the value of accuracy, specificity and sensitivity is better than the first analysis when considering one time-point.

Next, we perform an analysis with representation methods.

5.3 Representation techniques

We compare time-series representation methods, such as DWT, ESiG, and SAX. We proposed using these three methods to find the best representation of data and the representation that improve the classifier's capacity to distinguish the different stage of dementia. The first results in figure 5.4 show the comparison between the three representation methods used in this work.

The results in figure 5.4 do not demonstrate an improvement in the total score accuracy over summarization methods,and the results are similar to static data. In this dataset, there is a significant imbalance of classes (83/26). This imbalance remains in the test set, which means that accuracy may not be the best evaluation metric, but sensitivity and specificity. With this in mind, ESiG shows very promising results compared to the summary.

The DWT representation reveals promising techniques to lead with this type of data. With this method, as shown in figure 5.4, we obtain a mean of accuracies of 0.806, a sensitivity of 0.903, and specificity of 0.439. These results prove that this type of transformation possibly is more suitable for our data. On the other hand, the specificity shows that this transformation does not add much capacity to discover the converted case, only shows excellent results detecting the stable cases.

The results of ESiG representation are represented in figure 5.4. With the ESiG representation, we

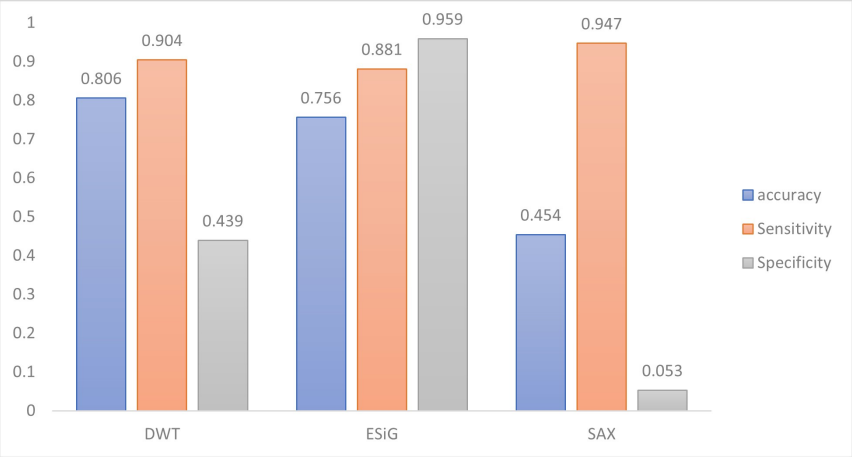

Figure 5.4: Classification results of representation methods.

obtain a mean accuracy value of 0.756, a sensitivity of 0.881, and a specificity of 0.959. We obtained excellent results in sensitivity and specificity. These results mean that the transformed ESiG representation results in a dataset in which the classifier can detect stable MCI and performs very well in the progression cases from MCI to AD, with a specificity of 0.959, which is the best result in our experiences.

The results shown in figure 5.4 by the SVM classifier with the SAX representation were the poor results in our experience. The values of mean accuracy of 0.454, a sensitivity of 0.947, and a specificity of 0.053. These results reflect the incapacity of classifier to deal with the representation of data using SAX. The initial transformation into the SAX word can lose the real meaning when transformed again into numerical values. That can be the reason for the poor results. So this method can not be considered to deal with this dataset for early detection of AD with this classifier.

Although the data representation methods have shown promise in previous studies, like SAX, these results may show us that this dataset may not contain enough information, contain too many variables to consider, or have many missing values. One of the big problems with real data is the number of missing values, which makes it impossible to interpret the data in the way we want.

5.4 Symbolic aggregate approximation and Vector Space Model

To improve our results, we tried a recent method that combines data representation by SAX and classification, the SAX-VSM. The method for classification use rating the similarity of the previous representation with the SAX method. The results with these methods and comparison with the other representation techniques are represented in figure 5.5.

Figure 5.5 shows that the SAX-VSM method improves, in terms of accuracy and specificity, SAX representation in SVM classifier but remains inferior compared with previous analysis. This approach has an accuracy of 0.6, a sensitivity of 0.73, and a specificity of 0.2. The accuracy and specificity results

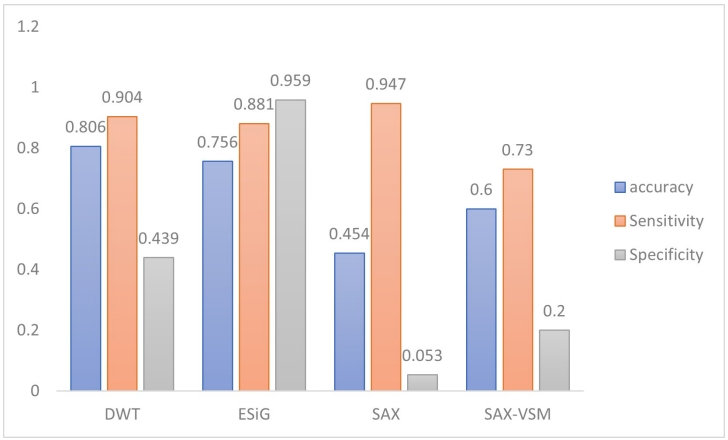

Figure 5.5: Representation methods results adding SAX-VSM.

are very low, which means that this representation with this data can not be used for our goal.

We need to make some changes in our approach to improve our results. In this sense, we try to fill the missing values with a different approach.

5.5 Last observation carried forward

To improve the results obtained, we decide to experiment a different approach to deal with a missing value. Instead of replacing all missing values with each column's mean, we decide to consider the assumption that the missing value is replaced by the value of the last observation or by zero in case we do not have any value observed. Figure 5.6 shows the results obtained.

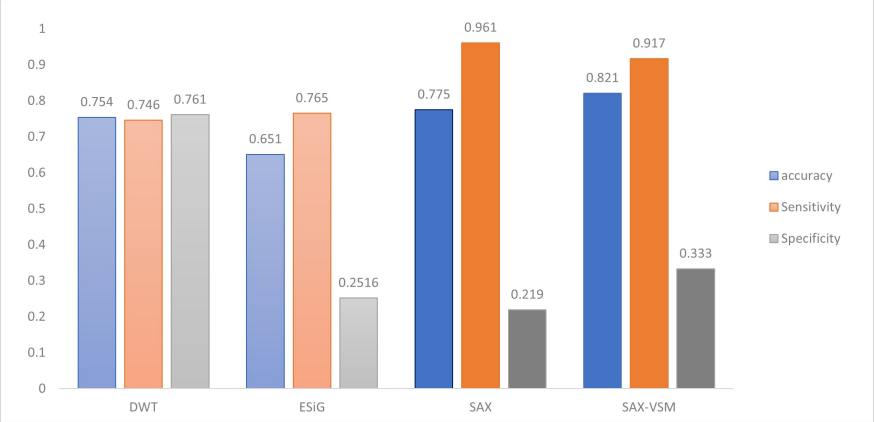

Figure 5.6: Classification results with last observation carried forward.

Figure 5.6 shows that the DWT and ESiG transformation do not improve but SAX and SAX-VSM

had better results than presented in figure 5.6. Although accuracy has slightly worsened in the DWT case, the increase in specificity value is worth it, which results in a better classifier. So, the lowest result obtained earlier is improved with this new dataset constructed with the last observation carried forward approach.

Figure 5.6 shows an improvement of accuracy value of SAX representation of 0.755, versus 0.454, and an improvement on SAX-VSM accuracy of 0.821, versus 0.6. Similarly, the sensitivity is improved in the case of SAX-VSM, from 0.73 to 0.917. In these two cases, the specificity remains very low but better than the initial results with SAX and SAX-VSM, 0.219 and 0.333, respectively.

5.6 Comparison/discussion

When all of these experiences are compared together, we can observe that with the temporal transformation performed in this work, we improved classification results from static data with one appointment to temporal data considering three follow-up evaluations and time-series representation.

In figure 5.7 we can observe the comparison of accuracy values achieve between all techniques used throughout this work. Regarding the accuracy metrics, we can observe an improvement, represented by dark blue in figure 5.7, in one summarization technique, median, and two representation techniques, DWT and SAX-VSM. However, in the other cases, we do not observe a big difference between the static data and other techniques.

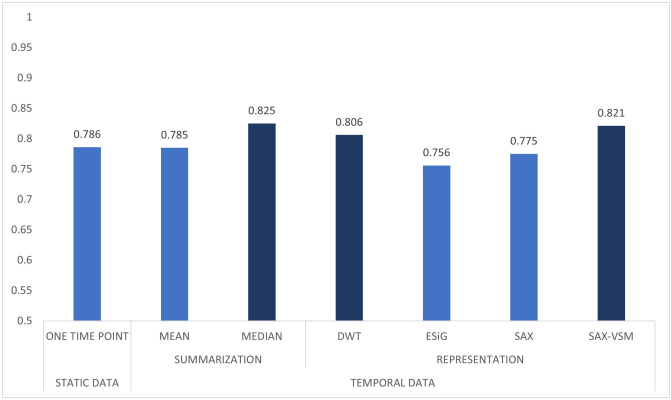

Figure 5.7: Comparison of accuracys metrics used to evaluated classsifiers performance.

Relatively to the comparison of sensibility obtained throughout this work, that represents the ability of classifiers detect stables MCI, all experience results into an improvement of this ability, as shown in figure 5.8.

The figure 5.8 shows improvements between sensibility achieve by static data and any other technique used. This improvement is significant and more relevant in representation techniques using SAX.

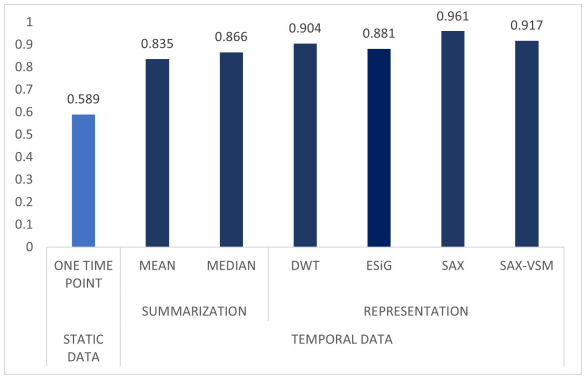

Figure 5.8: Comparison of sensitivity metrics used to evaluated classifiers performance.

The last comparison performed is between the specificity obtained. Figure 5.9 shows that four approaches have the ability to distinguish the minority class, that in our case is the important class which represents the converted case to AD.

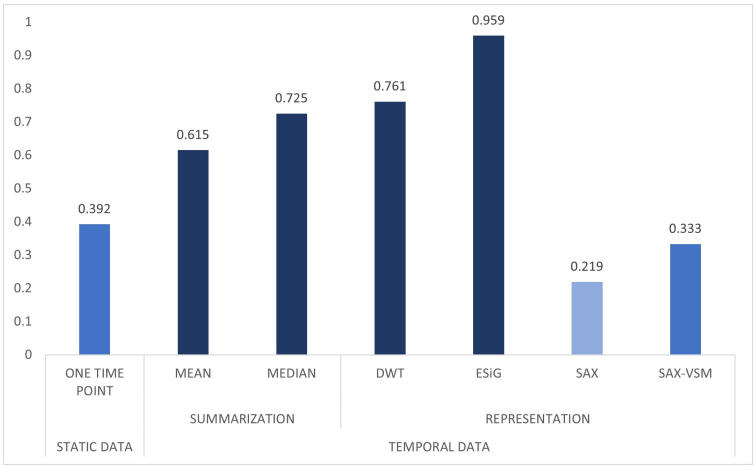

Figure 5.9: Comparison of specificity metrics used to evaluated classsifiers performance.

Figure 5.9 demonstrates that the summarization and two representation techniques used during this work improve specificity values. On the other hand, the SAX representation results remain at the level of static data in terms of specificity.

The improvement of the values in SAX representation does not happen in this case, and that means it may not be the best option to deal with our dataset and our problem view to predicting the progression from MCI into AD.

After all these comparisons, we can assume that the summarization techniques and some represen-

tation techniques can improve the results on the prediction of conversion of MCI to AD. In particular, ESiG may not have the higher accuracy but it seems to be a robust approach to handle this kind of data, since it presents higher sensitivity and specificity values. Additionally, we can also conclude that the way we treat the missing values directly interfered with the classifier's final result. In this sense, our experience demonstrates that the last observation carried forward does not impact the same way in different representation techniques.

5.7 summary

In this chapter, we used two distinct approaches, using static data and temporal data, to predict conversion from MCI to AD. To define the class labels, we use a time window approach in these two types of data. Initially, we use static data, with information about neuropsychological tests from one medical appointment, to select the classifier to be used in the next steps. Next, we use temporal data with information about three medical appointments and a class label defined by a three-year time-window approach with SVM, which was selected in static analysis. With these data, we used more detailed information about the evolution of the disease. The results revealed that different summarization techniques and some representation increases the prediction ability of the models.

Chapter 6

Conclusion

In this work, was studied the diagnosis and prognosis prediction for patients with Alzheimer's Disease. Therefore was performed a comparison between static data, only one appointment, to temporal data, using three appointments information. A dataset consisting of neuropsychological tests, demographics, and the corresponding diagnosis was considered to perform this work. Furthermore, a time-window approach to attribute the class label to instances was used. In terms of time-windows, For static data are considered four time-windows, from two to five years.

Regarding temporal data and considering the data available, was chosen a time-window of three years. Because it was added to the dataset along the time new information about neuropsychological tests, this dataset has missing values. There are many more MCI instances in the dataset than AD instances, resulting in class imbalance. In the case of an imbalanced class dataset, the sensitivity and specificity metrics are used in all models to evaluate the results without bias.

In this work, was created a model that can distinguish the patients with stable MCI from those probably converted in AD. For better understanding and develop of the model was analyzed a set of supervised data mining algorithms, such as Naive-Bayes (NB), Decision-Tree (DT), Support Vector Machine (SVM), and Neural-Networks (NN). The conclusion was that the SVM has a better performance when using static data (only one time point). For this reason, this algorithm was chosen for the remaining analysis. In the temporal data, an approach to predicting the conversion from MCI to AD that uses information from three appointments instead of using the first and last evaluation of the patient was used. To define the different profiles, it was applied a time-window approach after the third appointment. The time-window with the highest discriminative power is the three-year time-window. The main goal is to find the best technique to extract useful information with temporal data by comparing the summarization techniques with the representation techniques and analyze his improvements compared to static data analysis. One of the best diagnostic models were obtained using the SVM algorithm and the ESiG representation, with an accuracy of 0.756, a sensitivity of 0.881 a specificity of 0.959. The bests results achieve by applying summarization techniques are with the median, with an accuracy 0.825, a sensitivity 0.866, and a specificity of 0.725. These results are improved from static data with the representation

techniques, especially with the EsiG representation. In the overall analysis, it is possible to conclude that using three time-points is better than only one time-point in predicting conversion from MCI to AD, and the representation techniques improve the results compared with static data and summarization techniques.

6.1 Future Work

Some other techniques should be tried to deal with missing values in future work, which are significantly related to results and considering the amount of missing value that this dataset has. For instance, techniques based on deep learning for missing value imputation should be tested in this dataset to evaluate its impact in final performance. Furthermore, it is essential to describe the different groups of patients using clustering techniques, for example, to group converted patients and not converted patients. This approach can allow the development of specialized models for specific groups of patients and improve prediction accuracy. Finally, if all this information is used and applied to medical assistance, it can significantly evolve in precise and earlier diagnosis in AD and is possibly used daily by doctors.

Bibliography

[1] Stanisław Adaszewski et al. “How early can we predict Alzheimer’s disease using computational anatomy?” In: *Neurobiology of aging* 34.12 (2013), pp. 2815–2826.

[2] Jack Albright, Alzheimer’s Disease Neuroimaging Initiative, et al. “Forecasting the progression of Alzheimer’s disease using neural networks and a novel preprocessing algorithm”. In: *Alzheimer’s & Dementia: Translational Research & Clinical Interventions* 5 (2019), pp. 483–491.

[3] Association Alzheimer’s. “2015 Alzheimer’s disease facts and figures.” In: *Alzheimer’s & dementia: the journal of the Alzheimer’s Association* 11.3 (2015), p. 332.

[4] Imanol Perez Arribas. “Derivatives pricing using signature payoffs”. In: *arXiv preprint arXiv:1809.09466* (2018).

[5] American Psychiatric Association. *Diagnostic criteria from dsM-iV-tr*. American Psychiatric Pub, 2000.

[6] Nikhil Bhagwat et al. “Modeling and prediction of clinical symptom trajectories in Alzheimer’s disease using longitudinal data”. In: *PLoS computational biology* 14.9 (2018), e1006376.

[7] R Nick Bryan. *Machine learning applied to Alzheimer disease*. 2016.

[8] Nitesh V Chawla et al. “SMOTE: synthetic minority over-sampling technique”. In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[9] Ilya Chevyrev, Vidit Nanda, and Harald Oberhauser. “Persistence paths and signature features in topological data analysis”. In: *IEEE transactions on pattern analysis and machine intelligence* 42.1 (2018), pp. 192–202.

[10] DG Clark et al. “Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease”. In: *Cortex* 55 (2014), pp. 202–218.

[11] Yue Cui et al. “Identification of conversion from mild cognitive impairment to Alzheimer’s disease using multivariate predictors”. In: *PloS one* 6.7 (2011), e21896.

[12] Bradford C Dickerson et al. “Clinical prediction of Alzheimer disease dementia across the spectrum of mild cognitive impairment”. In: *Archives of General Psychiatry* 64.12 (2007), pp. 1443–1450.

[13] Carl Eckerström et al. “A combination of neuropsychological, neuroimaging, and cerebrospinal fluid markers predicts conversion from mild cognitive impairment to dementia”. In: *Journal of Alzheimer's disease* 36.3 (2013), pp. 421–431.

[14] Soheil Esmaeilzadeh et al. “End-to-end Alzheimer's disease diagnosis and biomarker identification”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2018, pp. 337–345.

[15] Carlos Alberto Barata Dias Garcia. “A Doença de Alzheimer: problemas do diagnóstico clınico”. In: (1984).

[16] Joseph Gaugler et al. “2019 Alzheimer's disease facts and figures”. In: *Alzheimers & Dementia* 15.3 (2019), pp. 321–387.

[17] Massimiliano Grassi et al. “A clinically-translatable machine learning algorithm for the prediction of alzheimer's disease conversion in individuals with mild and premild cognitive impairment”. In: *Journal of Alzheimer's Disease* 61.4 (2018), pp. 1555–1573.

[18] Massimiliano Grassi et al. “A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion: further evidence of its accuracy via a transfer learning approach”. In: *International psychogeriatrics* 31.7 (2019), pp. 937–945.

[19] Massimiliano Grassi et al. “A Novel Ensemble-Based Machine Learning Algorithm To Predict The Conversion From Mild Cognitive Impairment To Alzheimer's Disease Using Socio-demographic Characteristics, Clinical Information And Neuropsychological Measures”. In: *bioRxiv* (2019), p. 564716.

[20] Isabelle Guyon and André Elisseeff. “An introduction to variable and feature selection”. In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[21] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[22] John Hardy and Alain Israël. “Alzheimer's disease: in search of γ-secretase”. In: *Nature* 398.6727 (1999), p. 466.

[23] A Kormilitzin et al. “Named entity recognition in electronic health records using transfer learning bootstrapped neural networks”. In: *Neural Networks* 121 (2019).

[24] Andrey Kormilitzin et al. “Detecting early signs of depressive and manic episodes in patients with bipolar disorder using the signature-based model”. In: *arXiv preprint arXiv:1708.01206* (2017).

[25] Marcin Kruczyk et al. “Monte Carlo feature selection and rule-based models to predict Alzheimer's disease in mild cognitive impairment”. In: *Journal of neural transmission* 119.7 (2012), pp. 821–831.

[26] Sei J Lee et al. “A clinical index to predict progression from mild cognitive impairment to dementia due to Alzheimer's disease”. In: *PloS one* 9.12 (2014), e113535.

[27] Kan Li et al. “Prediction of conversion to Alzheimer’s disease with longitudinal measures and time-to-event data”. In: *Journal of Alzheimer’s Disease* 58.2 (2017), pp. 361–371.

[28] Jessica Lin et al. “Experiencing SAX: a novel symbolic representation of time series”. In: *Data Mining and knowledge discovery* 15.2 (2007), pp. 107–144.

[29] Pranav Patel Eamonn Keogh Jessica Lin and Stefano Lonardi. “Mining motifs in massive time series databases”. In: *IEEE Int. Conf. on Data Mining*. 2002.

[30] Xiaojing Long et al. “Prediction and classification of Alzheimer disease based on quantification of MRI deformation”. In: *PloS one* 12.3 (2017), e0173372.

[31] Joao Maroco et al. “Prediction of dementia patients: a comparative approach using parametric vs. non parametric classifiers”. In: *XVII Congresso Anual da Sociedade Portuguesa de Estatıstica*. Sociedade Portuguesa de Estatıstica. 2012, pp. 241–251.

[32] João Maroco et al. “Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests”. In: *BMC research notes* 4.1 (2011), p. 299.

[33] José Eduardo Martinelli et al. “Comparison of the diagnostic accuracy of neuropsychological tests in differentiating Alzheimer’s disease from mild cognitive impairment: can the montreal cognitive assessment be better than the Cambridge cognitive examination”. In: *Dementia and geriatric cognitive disorders extra* 4.2 (2014), pp. 113–121.

[34] PS Mathuranath et al. “A brief cognitive test battery to differentiate Alzheimer’s disease and frontotemporal dementia”. In: *Neurology* 55.11 (2000), pp. 1613–1620.

[35] Jordi A Matias-Guiu et al. “Comparative diagnostic accuracy of the ACE-III, MIS, MMSE, MoCA, and RUDAS for screening of Alzheimer disease”. In: *Dementia and geriatric cognitive disorders* 43.5-6 (2017), pp. 237–246.

[36] Theophano Mitsa. *Temporal data mining*. CRC Press, 2010.

[37] James Morrill et al. “The signature-based model for early detection of sepsis from electronic health records in the intensive care unit”. In: *2019 Computing in Cardiology (CinC)*. IEEE. 2019, Page–1.

[38] Telma Pereira et al. “Predicting progression of mild cognitive impairment to dementia using neuropsychological data: a supervised learning approach using time windows”. In: *BMC medical informatics and decision making* 17.1 (2017), p. 110.

[39] Ivan Popivanov and Renee J Miller. “Similarity search over time-series data using wavelets”. In: *Proceedings 18th international conference on data engineering*. IEEE. 2002, pp. 212–221.

[40] Florence Portet et al. “Mild cognitive impairment (MCI) in medical practice: a critical review of the concept and new diagnostic procedure. Report of the MCI Working Group of the European Consortium on Alzheimer’s Disease”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 77.6 (2006), pp. 714–718.

[41] Marialuisa Restaino et al. “Predicting risk of 2-year incident dementia using the CAMCOG total and subscale scores”. In: *Age and ageing* 42.5 (2013), pp. 649–653.

[42] Giovanna Ricci. “Social Aspects of Dementia Prevention from a Worldwide to National Perspective: A Review on the International Situation and the Example of Italy”. In: *Behavioural neurology* 2019 (2019).

[43] Pavel Senin and Sergey Malinchik. “Sax-vsm: Interpretable time series classification using sax and vector space model”. In: *2013 IEEE 13th international conference on data mining*. IEEE. 2013, pp. 1175–1180.

[44] William Rodman Shankle et al. “Improving dementia screening tests with machine learning methods”. In: *Alzheimer’s Research* 2.3 (1996).

[45] Ting Shen et al. “Decision Supporting Model for One-year Conversion Probability from MCI to AD using CNN and SVM”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 738–741.

[46] Dina Silva et al. “Prediction of long-term (5 years) conversion to dementia using neuropsychological tests in a memory clinic setting”. In: *Journal of Alzheimer’s Disease* 34.3 (2013), pp. 681–689.

[47] Matthias H Tabert et al. “Neuropsychological prediction of conversion to Alzheimer disease in patients with mild cognitive impairment”. In: *Archives of general psychiatry* 63.8 (2006), pp. 916–924.

[48] Jiliang Tang, Salem Alelyani, and Huan Liu. “Feature selection for classification: A review”. In: *Data classification: Algorithms and applications* (2014), p. 37.

[49] Maileen Gloriane Ulep, Simrit Kaur Saraon, and Samantha McLea. “Alzheimer disease”. In: *The Journal for Nurse Practitioners* 14.3 (2018), pp. 129–135.

[50] Alyssa Weakley et al. “Neuropsychological test selection for cognitive impairment classification: a machine learning approach”. In: *Journal of clinical and experimental neuropsychology* 37.9 (2015), pp. 899–916.

[51] Daoqiang Zhang et al. “Multimodal classification of Alzheimer’s disease and mild cognitive impairment”. In: *Neuroimage* 55.3 (2011), pp. 856–867.

[52] Yingying Zhu and Mert R Sabuncu. “A Probabilistic Disease Progression Model for Predicting Future Clinical Outcome”. In: *arXiv preprint arXiv:1803.05011* (2018).